

# Self-organization of microcircuits in networks of spiking neurons with plastic synapses

Gabriel Koch Ocker<sup>1,3</sup>, Ashok Litwin-Kumar<sup>2,3,4</sup>, Brent Doiron<sup>2,3\*</sup>

1: Department of Neuroscience, University of Pittsburgh, Pittsburgh, PA, United States of America

2: Department of Mathematics, University of Pittsburgh, Pittsburgh, PA, United States of America

3: Center for the Neural Basis of Cognition, University of Pittsburgh and Carnegie Mellon University, Pittsburgh, PA, United States of America

4: Center for Theoretical Neuroscience, Columbia University, New York, NY, United States of America

\* E-mail: bdoiron@pitt.edu

## Abstract

The synaptic connectivity of cortical networks features an overrepresentation of certain wiring motifs compared to simple random-network models. This structure is shaped, in part, by synaptic plasticity that promotes or suppresses connections between neurons depending on their spiking activity. Frequently, theoretical studies focus on how feedforward inputs drive plasticity to create this network structure. We study the complementary scenario of self-organized structure in a recurrent network, with spike timing-dependent plasticity driven by spontaneous dynamics. We develop a self-consistent theory that describes the evolution of network structure by combining fast spiking covariance with a fast-slow theory for synaptic weight dynamics. Through a finite-size expansion of network dynamics, we obtain a low-dimensional set of nonlinear differential equations for the evolution of two-synapse connectivity motifs. With this theory in hand, we explore how the form of the plasticity rule drives the evolution of microcircuits in cortical networks. When potentiation and depression are in approximate balance, synaptic dynamics depend on the frequency of weighted divergent, convergent, and chain motifs. For additive, Hebbian STDP, these motif interactions create instabilities in synaptic dynamics that either promote or suppress the initial network structure. Our work provides a consistent theoretical framework for studying how spiking activity in recurrent networks interacts with synaptic plasticity to determine network structure.

## Author Summary

The connectivity of mammalian brains exhibits structure at a wide variety of spatial scales, from the broad (which brain areas connect to which) to the extremely fine (where synapses from different inputs lie on the morphology of individual neurons). Recent experimental work in the neocortex has highlighted structure at the level of microcircuits: different patterns of connectivity between small groups of neurons are either more or less abundant than would be expected by chance. A central question in systems neuroscience is how this structure emerges. Attempts to answer this question are confounded by the known mutual interaction of network structure and spiking activity. Indeed, synaptic connections influence spiking statistics, while individual synapses are highly plastic and become stronger or weaker depending on the activity of the pre- and postsynaptic neurons. We present a self-consistent theory for how activity-dependent synaptic plasticity leads to the emergence of neuronal microcircuits and use it to show how the form of the plasticity rule can govern the promotion or suppression of different connectivity patterns. Our work provides a foundation for understanding how cortical circuits, and not just individual synapses, are malleable in response to inputs both external and internal to a network.

## Introduction

The wiring of neuronal networks exhibits structure across a broad range of spatial scales [1]. For example, patterns of connectivity among small groups of cortical neurons are over- or under-represented compared to random networks [2–5]. The prevalence of these motifs is related to a neuron’s stimulus preferences and activity levels [6, 7]. Motivated in part by these observations, there is a growing body of theoretical work that discusses how wiring structure dictates the coordinated spiking activity of cortical neurons in recurrent networks [8–18].

While neural architecture undoubtedly plays a strong role in determining neuronal activity, the reverse is also true. Individual synapses can both potentiate (strengthen) and depress (weaken), and whether they do so depends on the relative timing of action potentials from the connected neurons [19, 20]. Such *spike timing-dependent plasticity* (STDP) has featured prominently in both experimental and theoretical studies of neural circuits [21–23]. Of particular interest, STDP provides a mechanism for Hebbian plasticity: synaptic potentiation occurs when a presynaptic neuron reliably drives spike responses from a postsynaptic neuron, while failure to recruit spiking results in synaptic depression [24]. Hebbian plasticity provides a potential link between circuit structure and function by providing a mechanism for the formation of heavily wired assemblies of neurons, where assembly membership is associated with coordinated, elevated firing rates during a specific computation [25]. Evidence supporting this idea, originally proposed by Hebb [26], has been found in both hippocampus [27] and sensory cortex [28].

Despite the promise of STDP to provide insight into the functional wiring of large neural circuits, many studies of STDP have concentrated on the plasticity of synaptic connections between just a single pair of pre- and postsynaptic neurons, often focusing on the distribution of individual synaptic weights [24, 29–32]. Other studies have shown that multiple temporally correlated inputs to a neuron will cooperate to potentiate, while uncorrelated inputs may depress [24, 33–35]. In this case STDP can generate feedforward circuits [36], which while important for the propagation of neural activity [37], are unlike the recurrent structure of the neocortex. Understanding the two-way interaction between plastic recurrent network structure and spiking activity is thus a central challenge for theories of synaptic plasticity.

Due to this challenge, many studies have resorted to large scale numerical simulations of cortical networks with plastic synapses [38–41]. While intuition for the development of circuit structure can be gained using this approach, without a governing theoretical framework it is often difficult to extract generalized principles. Alternatively, mathematical analyses have been restricted to either small networks [40, 42], or have required the assumption that neurons fire as Poisson processes [43–46]. These latter works assumed shared inputs from outside the network to be the only source of correlated spiking activity, neglecting covariance due to recurrent coupling. Thus, there is a need for a coherent mathematical framework that captures how STDP drives self-organization of circuit structure in recurrent cortical networks.

To this end, we construct a self-consistent theory for the coevolution of spiking statistics and synaptic weights in networks with STDP. This theory makes use of a previously developed linear response framework for calculating joint spiking statistics [15, 47, 48] and a separation of timescales between spiking covariance and synaptic plasticity [33]. Most previous studies of plasticity in recurrent networks have focused on how they can be trained to represent an external stimulus. We focus on how spiking covariance generated by coupling within the network interacts with plasticity to shape network structure. We then use this high-dimensional theory to derive a low-dimensional, closed system for STDP of synaptic connectivity motifs in recurrent networks. This reveals instabilities in the motif dynamics such that when potentiation and depression are approximately balanced, the dynamics are partitioned into regimes in which different motifs are promoted or suppressed depending on the initial network structure. It also highlights the

circumstances in which spike time covariations, in contrast to firing rates, drive STDP. In total, we provide a consistent and general framework in which to study STDP in large recurrent networks.

## Results

Our study is separated into two main sections. The first presents a self-consistent theory for STDP in recurrent networks of model spiking neurons. Using this theory, we accurately predict the co-evolution of spiking covariance on fast timescales with synaptic weights on slow timescales of a large network. The second part leverages our theory to develop a low-dimensional dynamical system for the development of two-synapse motifs in the network structure. We analyze this system and determine how the balance between synaptic potentiation and depression drives the emergence of microcircuits in recurrent networks.

### Spike train covariance determines synaptic plasticity

We begin by reviewing a well studied phenomenological model of spike timing-dependent plasticity (STDP) [49], acting within a simple circuit of two reciprocally coupled neurons. Consider a pair of pre- and postsynaptic spike times with time lag  $s = t_{\text{post}} - t_{\text{pre}}$ . The evolution of the synaptic weight connecting presynaptic neuron  $j$  to postsynaptic neuron  $i$  obeys  $\mathbf{W}_{ij} \rightarrow \mathbf{W}_{ij} + L(s)$ , with the STDP rule  $L(s)$  (Fig. 1A) being Hebbian:

$$L(s) = \begin{cases} \mathcal{H}(W^{\max} - \mathbf{W}_{ij}) f_+ e^{-\frac{|s|}{\tau_+}}, & \text{if } s \geq 0 \\ \mathcal{H}(\mathbf{W}_{ij}) (-f_-) e^{-\frac{|s|}{\tau_-}}, & \text{if } s \leq 0, \end{cases} \quad (1)$$

Here  $\mathcal{H}(x) = 1$  if  $x > 0$  while  $\mathcal{H}(x) = 0$  if  $x < 0$ , imposing bounds on the weights to prevent the magnitude of excitatory synapses from becoming negative or potentiating without bound (i.e.  $0 \leq \mathbf{W}_{ij} \leq W^{\max}$ ). The coefficients  $f_{\pm}$  scale the amplitude of weight changes induced by individual pre-post spike pairs and  $\tau_{\pm}$  determine how synchronous pre- and postsynaptic spikes must be to drive plasticity.

The spike train from neuron  $i$  is the point process  $\mathbf{y}_i(t) = \sum_k \delta(t - t_{ik})$ , with  $t_{ik}$  being its  $k^{\text{th}}$  spike time. Following [33] we relate the joint statistics of  $\mathbf{y}_i(t)$  and  $\mathbf{y}_j(t)$  to the evolution of synaptic weights. We first assume that individual pre-post spike pairs induce small changes in synaptic weights ( $f_{\pm} \ll W^{\max}$ ). This makes synaptic weights evolve slowly, on a much longer timescale than the millisecond scale of pairwise spiking covariance due to network interactions. The separation of timescales between synaptic plasticity and spiking activity provides an approximation to the evolution of the synaptic weights (Methods: learning dynamics):

$$\frac{d\mathbf{W}_{ij}}{dt} = \mathbf{W}_{ij}^0 \int_{-\infty}^{\infty} L(s) (r_i r_j + \mathbf{C}_{ij}(s)) ds. \quad (2)$$

Here  $r_i = \langle \mathbf{y}_i(t) \rangle$  is the time-averaged firing rate of neuron  $i$ , and  $\mathbf{C}_{ij}(s) = \langle (\mathbf{y}_i(t) - r_i)(\mathbf{y}_j(t+s) - r_j) \rangle$  is the cross-covariance function of neuron  $i$  and  $j$ 's spike trains. The separation of timescales allows us to calculate the equilibrium spiking statistics  $\mathbf{C}$ , taking  $\mathbf{W}$  to be constant on the timescale of  $\mathbf{C}(s)$ . The term  $r_i r_j$  in Eq. (2) captures the firing rate dependence of STDP, while  $\mathbf{C}_{ij}(s)$  models the sensitivity of STDP to spike timing. Finally,  $\mathbf{W}^0$  is the adjacency matrix of the network – a binary matrix with  $\mathbf{W}_{ij}^0 = 1$  denoting the presence of a synapse. Multiplying by  $\mathbf{W}_{ij}^0$  ensures that synapses that do not exist cannot potentiate into existence. Eq. (2) requires only the first and second order joint spiking statistics. To facilitate calculations, many previous studies have used Poisson neuron models with a specified  $r_i$  and  $\mathbf{C}_{ij}(s)$  to generate  $\mathbf{y}_i(t)$ . In contrast, we will use a white noise driven exponential integrate-and-fire model [50] for the generation of spike times (Methods: Network model). While this complicates the calculation of the spike train statistics, it provides a more biophysically realistic model of neural dynamics [51, 52] that better captures the timescales and neuronal nonlinearities that shape  $r_i$  and

$\mathbf{C}_{ij}(s)$ . In total, the above theory determines synaptic evolution from the integrated combination of an STDP rule  $L(s)$  and the spike train cross-covariance function  $\mathbf{C}_{ij}(s)$ . Thus, any mechanism affecting two neurons' spiking covariance is expected to shape network structure through STDP.

As a simple illustration of how spiking correlations can drive STDP, we examined the synaptic weight dynamics,  $\mathbf{W}_{12}(t)$  and  $\mathbf{W}_{21}(t)$ , in a reciprocally coupled pair of neurons, both in the presence and absence of common inputs. Specifically, the fluctuating input to neuron  $i$  was the sum of a private and common term,  $\sqrt{1-c}\xi_i(t) + \sqrt{c}\xi_c(t)$ , with  $c$  the fraction of shared input to the neurons. In the absence of common input ( $c = 0$ ; Fig. 1B), the two synapses behaved as expected with Hebbian STDP: one synapse potentiated and the other depressed (Fig. 1C). In contrast, the presence of common input ( $c = 0.05$ ) was a source of synchrony in the two neurons' spike trains, inducing a central peak in the spike train cross-covariance function  $\mathbf{C}_{ij}(s)$  (Fig. 1B vs 1D). In this case, both synapses potentiated (Fig. 1E) because the common input increased synchronous spiking. Since the potentiation side of the learning rule was sharper than the depression side (Fig. 1A), this enhanced the degree of overlap between  $\mathbf{C}_{ij}(s)$  and the potentiation component of  $L(s)$ . This overcame the effects of depression in the initially weaker synapse and promoted strong, bidirectional connectivity in the two-neuron circuit.

This example highlights how the temporal shape of the spike train cross-covariance function can interact with the shape of the learning rule,  $L(s)$ , to direct spike timing-dependent plasticity. However, this case only considered the effect of correlated inputs from outside of the modeled circuit (Fig. 1). Our primary goal is to predict how spiking covariance due to internal network interactions combines with STDP to drive self-organized network structure. In order to do this, we first require a theory for predicting the spiking covariance between all neuron pairs given some static, recurrent connectivity. Once this theory has been developed, we will use it to study the case of plastic connectivity.

## Network architecture determines spiking covariance in static networks

In this section we review approximation methods [15, 47, 48] that estimate the pairwise spike train cross-covariances  $\mathbf{C}_{ij}(s)$  using a static weight matrix  $\mathbf{W}$  (see Methods: Spiking statistics for a full description). The exposition is simplified if we consider the Fourier transform of a spike train,  $\mathbf{y}_i(\omega) = \int_{-\infty}^{\infty} \mathbf{y}_i(t) e^{-2\pi i \omega t} dt$ , where  $\omega$  is frequency. Assuming weak synaptic connections  $\mathbf{W}_{ij}$ , we approximate the spike response from neuron  $i$  as:

$$\mathbf{y}_i(\omega) = \mathbf{y}_i^0(\omega) + \mathbf{A}_i(\omega) \left( \sum_{j=1}^N \mathbf{W}_{ij} J(\omega) \mathbf{y}_j(\omega) \right). \quad (3)$$

The function  $\mathbf{A}_i(\omega)$  is the linear response [53] of the postsynaptic neuron, measuring how strongly modulations in synaptic currents at frequency  $\omega$  are transferred into modulations of instantaneous firing rate about a background state  $\mathbf{y}_i^0$ . The function  $J(\omega)$  is a synaptic filter. In brief, Eq. (3) is a linear ansatz for how a neuron integrates and transforms a realization of synaptic input into a spike train.

Following [15, 47, 48] we use this linear approximation to estimate the Fourier transform of  $\mathbf{C}_{ij}(s)$ , written as  $\mathbf{C}_{ij}(\omega) = \langle \mathbf{y}_i(\omega) \mathbf{y}_j^*(\omega) \rangle$ ; here  $\mathbf{y}^*$  denotes complex conjugation. This yields the following matrix equation:

$$\mathbf{C}(\omega) = \left( \mathbf{I} - (\mathbf{W} \cdot \mathbf{K}(\omega)) \right)^{-1} \mathbf{C}^0(\omega) \left( \mathbf{I} - (\mathbf{W} \cdot \mathbf{K}^*(\omega)) \right)^{-1}, \quad (4)$$

where  $\mathbf{K}(\omega)$  is an interaction matrix defined by  $\mathbf{K}_{ij}(\omega) = \mathbf{A}_i(\omega) \mathbf{J}_{ij}(\omega)$ . The matrix  $\mathbf{C}^0(\omega)$  is the baseline covariance, with elements  $\mathbf{C}_{ij}^0(\omega) = \langle \mathbf{y}_i^0(\omega) \mathbf{y}_j^{0*}(\omega) \rangle$ , and  $\mathbf{I}$  is the identity matrix. Using Eq. (4) we recover the matrix of spike train cross-covariance functions  $\mathbf{C}(s)$  by inverse Fourier transformation. Thus, Eq. (4)

provides an estimate of the statistics of pairwise spiking activity in the full network, taking into account the network structure.

As a demonstration of the theory, we examined the spiking covariances of three neurons from a 1,000-neuron network (Fig. 2A, colored neurons). The synaptic weight matrix  $\mathbf{W}$  was static and had an adjacency matrix  $\mathbf{W}^0$  that was randomly generated with Erdős-Rényi statistics (connection probability of 0.15). The neurons received no correlated input from outside the network, making  $\mathbf{C}^0(\omega)$  a diagonal matrix, and thus recurrent network interactions were the only source of spiking covariance. Neuron pairs that connected reciprocally with equal synaptic weights had temporally symmetric spike train cross-covariance functions (Fig. 2C), while uni-directional connections gave rise to temporally asymmetric cross-covariances (Fig. 2D). When neurons were not directly connected, their covariance was weaker than that of directly connected neurons but was still nonzero (Fig. 2E). The theoretical estimate provided by Eq. (4) was in good agreement with estimates from direct simulations of the network (Fig. 2C,D,E red vs. gray curves).

## Self-consistent theory for network structure and spiking covariance with plastic synapses

In general, it is challenging to develop theoretical techniques for stochastic systems with several variables and nonlinear coupling [53], such as in Eq. (2). Fortunately, in our model the timescale of spiking covariance in the recurrent network with static synapses is on the order of milliseconds (Fig. 2C,D,E), while the timescale of plasticity is minutes (Fig. 1C,E). This separation of timescales provides an opportunity for a self-consistent theory for the coevolution of  $\mathbf{C}(s)$  and  $\mathbf{W}(t)$ . That is, so long as  $f_{\pm}$  in Eq. (1) are sufficiently small, we can approximate  $\mathbf{W}$  as static over the timescales of  $\mathbf{C}(s)$  and insert Eq. (4) into Eq. (2). The resulting system yields a solution  $\mathbf{W}(t)$  that captures the long timescale dynamics of the plastic network structure (Methods: self-consistent theory for network plasticity).

As a first illustration of our theory, we focus on the evolution of three synaptic weights in a 1,000-neuron network (Fig. 3A, colored arrows). The combination of Eqs. (2) and (4) predicted the dynamics of  $\mathbf{W}(t)$ , whether the weight increased with time (Fig. 3B left, red curve), decreased with time (Fig. 3C left, red curve), or remained approximately constant (Fig. 3D left, red curve). In all three cases, the theory matched well the average evolution of the synaptic weight estimated from direct simulations of the spiking network (Fig. 3B,C,D left, thick black curves). Snapshots of the network at three time points (axis arrows in Fig. 3B,C,D, left), showed that  $\mathbf{W}$  coevolved with the spiking covariance (Fig. 3B,C,D right). We remark that for any realization of background input  $\mathbf{y}^0(t)$ , the synaptic weights  $\mathbf{W}(t)$  deviated from their average value with increasing spread (Fig. 3B,C,D left, thin black curves). This is expected since  $\mathbf{C}(t)$  is an average over realizations of  $\mathbf{y}^0(t)$ , and thus provides only a prediction for the drift of  $\mathbf{W}(t)$ , while the stochastic nature of spike times leads to diffusion of  $\mathbf{W}(t)$  around this drift [33].

In sum, the fast-slow decomposition of spiking covariance and synaptic plasticity provides a coherent theoretical framework to investigate the formation of network structure through STDP. Also, our treatment is complementary to past studies on STDP [35, 44, 54] that focused on the development of architecture through external input. In our model, there is no spiking covariance in the background state (i.e.  $\mathbf{C}_{ij}^0(s) = \langle \mathbf{y}_i^0(t+s) \mathbf{y}_j^0(t) \rangle = 0$  for  $i \neq j$ ). Hence, we are specifically considering self-organization of network structure through internally generated spiking covariance.

While our theory gives an accurate description of plasticity in the network, it is nevertheless high-dimensional. Keeping track of every individual synaptic weight and spike train cross-covariance function involves  $\mathcal{O}(N^2)$  variables. For large networks, this becomes computationally challenging. More importantly,

this high-dimensional theory does not provide insights into the plasticity of the *connectivity patterns* or *motifs* that are observed in cortical networks [3, 4]. Motifs involving two or more neurons represent correlations in the network's weight matrix, which cannot be described by a straightforward application of mean-field techniques. In the next sections, we develop a principled approximation of the high-dimensional theory to a closed low-dimensional mean-field theory for how the mean weight and the strength of two-synapse motifs evolve due to STDP.

## Dynamics of mean synaptic weight

We begin by considering the simple case of a network with independent weights (an Erdős-Rényi network). In this case, the only dynamical variable characterizing network structure is the mean synaptic weight,  $p$ :

$$p = \frac{1}{N^2} \sum_{i,j} \mathbf{W}_{ij}. \quad (5)$$

The mean synaptic weight,  $p$ , in addition to the connection probability  $p_0$ , completely characterizes the connectivity of a weighted Erdős-Rényi network. In order to calculate the dynamics of  $p$ , we insert the fast-slow STDP theory of Eq. (2) into Eq. (5):

$$\frac{dp}{dt} = \frac{1}{N^2} \sum_{i,j} \mathbf{W}_{ij}^0 \int_{-\infty}^{\infty} L(s)(r_i r_j + \mathbf{C}_{ij}(s)) ds, \quad (6)$$

where the spiking covariances are calculated using linear response theory (Eq. (4)). This equation depends on the network structure in two ways. First, it depends on the full adjacency matrix  $\mathbf{W}^0$ . Second, the spike train cross-covariances depend on the full weight matrix:  $\mathbf{C}_{ij}(s) = \mathbf{C}_{ij}(s; \mathbf{W})$ . This dependence of a first-order connectivity statistic on the network structure poses a challenge for the development of a closed theory.

The main steps in our approach here are two approximations. First, the matrix of spike train cross-covariances  $\mathbf{C}(s)$  obtained from our linear ansatz (Eq. (4)) can be expanded in a power series around the background cross-covariances  $\mathbf{C}^0(s)$  (see Eq. (27)). Powers of the interaction matrix  $\mathbf{K}$  in this series correspond to different lengths of paths through the network [13, 15]. We truncate the spiking covariances at length one paths to obtain:

$$\mathbf{C}_{ij}(s) \approx \underbrace{(\mathbf{W}_{ij} \mathbf{K}_{ij} * \mathbf{C}_{jj}^0)(s)}_{\text{forward}} + \underbrace{(\mathbf{C}_{ii}^0 * \mathbf{W}_{ji} \mathbf{K}_{ji}^-)(s)}_{\text{backward}} + \underbrace{\sum_k (\mathbf{W}_{ik} \mathbf{K}_{ik} * \mathbf{C}_{kk}^0 * \mathbf{W}_{jk} \mathbf{K}_{jk}^-)(s)}_{\text{common}}, \quad (7)$$

where  $*$  denotes convolution. This truncation separates the sources of covariance between the spiking of neurons  $i$  and  $j$  into direct forward ( $i \leftarrow j$ ) and backward ( $i \rightarrow j$ ) connections, and common ( $k \rightarrow i$  and  $k \rightarrow j$ ) inputs. Nevertheless, after truncating  $\mathbf{C}(s)$ , the mean synaptic weight still depends on higher-order connectivity motifs (Eq. (32)). Fortunately, for an Erdős-Rényi network, these higher-order motifs are negligible.

The second approximation is to ignore the bounds on the synaptic weight in Eq. (1). While this results in a theory that only captures the transient dynamics of  $\mathbf{W}(t)$ , it greatly simplifies the derivation of the low-dimensional dynamics of motifs, because dynamics along the boundary surface are not considered.

With these two approximations, the mean synaptic weight obeys:

$$\frac{dp}{dt} = r^2 S \frac{1}{N^2} \sum_{i,j} \mathbf{W}_{ij}^0 + S_F \frac{1}{N^2} \sum_{i,j} \mathbf{W}_{ij}^0 \mathbf{W}_{ij} + S_B \frac{1}{N^2} \sum_{i,j} \mathbf{W}_{ij}^0 \mathbf{W}_{ji} + S_C \frac{1}{N^2} \sum_{i,j,k} \mathbf{W}_{ij}^0 \mathbf{W}_{ik} \mathbf{W}_{jk}. \quad (8)$$

The first term on the right hand side of Eq. (8) is scaled by  $S = \int_{-\infty}^{\infty} L(s)ds$ , modeling the interaction between STDP rules that lead to net potentiation ( $S > 0$ ) or depression ( $S < 0$ ), and the mean firing rate,  $r$ , across the network. This captures STDP due to chance spiking coincidence. The remaining terms capture how synaptic weights interact with the temporal structure of spiking covariance. Because of the expansion in Eq. (7), these dependencies decompose into three terms, each scaled by the integral of the product of the STDP rule  $L(s)$  and a component of the spike train cross-covariance  $\mathbf{C}(s)$ . Specifically, covariance due to forward connections is represented by  $S_F$  (Eq. (36); Fig. 4A), covariance due to backward connections is represented by  $S_B$  (Eq. (37); Fig. 4B), and finally covariance due to common connections is represented by  $S_C$  (Eq. (38); Fig. 4C).

For an Erdős-Rényi network, each sum in Eq. (8) can be simplified. Let  $p_0 = \frac{1}{N^2} \sum_{i,j} \mathbf{W}_{ij}^0$  be the connection probability of the network. Since our theory for spiking covariances required weak synapses, we also explicitly scaled the weights, motifs, and amplitude of synaptic changes  $f_{\pm}$  by  $\epsilon = 1/(Np_0)$ . This ensured that as the connection probability  $p_0$  was varied, synaptic weights scaled to keep the total input to a neuron constant (neglecting plasticity). The first and second terms of Eq. (8) correspond to the definitions of  $p_0$  and  $p$ . Since different elements of  $\mathbf{W}^0$  and  $\mathbf{W}$  are independent, the third term reduces to  $\frac{1}{N^2} \sum_{i,j} \mathbf{W}_{ij}^0 \mathbf{W}_{ji} = \epsilon p p_0 + \mathcal{O}(\epsilon^{3/2})$  due to the central limit theorem. The last term can be similarly evaluated and the dynamics of  $p$  to first order in  $\epsilon$  reduce to:

$$\frac{dp}{dt} = p_0 r^2 S + \epsilon (p(S_F + p_0 S_B) + p^2 S_C). \quad (9)$$

We next study this mean-field theory in two regimes, before examining the plasticity of non-Erdős-Rényi networks that exhibit motifs.

## Unbalanced STDP of the mean synaptic weight

Eq. (9) contains one term proportional to product of firing rates and the integral of the STDP rule,  $r^2 S$ , and additional terms proportional to the small parameter  $\epsilon$ . When the learning rule,  $L(s)$ , is dominated by either depression or potentiation (so that  $S \sim \mathcal{O}(1) \gg \epsilon$ ) the whole network uniformly depresses (Fig. 5A,C) or potentiates (Fig. 5B,D) due to chance spike coincidences (the firing rate term dominates in Eq. (2)). These dynamics are straightforward at the level of individual synapses and this intuition carries over straightforwardly to the mean synaptic weight. When the STDP rule is dominated by potentiation or depression, the  $\mathcal{O}(\epsilon)$  terms in Eq. (9) are negligible; the average plasticity is solely determined by the firing rates, with spiking covariance playing no role. In this case, the leading-order dynamics of  $p$  are:

$$p(t) = p_0 r^2 S t + p(0), \quad (10)$$

so that the mean synaptic weight either potentiates to its upper bound  $p_0 W^{\max}$  or depresses to 0, depending on whether the integral of the STDP rule,  $S$ , is positive or negative. For both depression and potentiation dominated STDP our simple theory in Eq. (10) quantitatively matches  $p(t)$  estimated from simulations of the entire network (Fig. 5C,D, black vs. red curves).

## Balanced STDP of the mean synaptic weight

On the other hand, if there is a balance between potentiation and depression in the STDP rule  $L(s)$ , then spiking covariance affects the average plasticity. In order to make explicit the balance between potentiation and depression, we write  $S = \pm \delta \epsilon$  (with  $+\delta \epsilon$  for STDP with the balance tilted in favor of potentiation and  $-\delta \epsilon$  for balance tilted in favor of depression). The leading-order dynamics of  $p$  are then, for Erdős-Rényi networks,

$$\frac{1}{\epsilon} \frac{dp}{dt} = \pm \delta p_0 r^2 + p(S_F + p_0 S_B) + p^2 S_C. \quad (11)$$

This quadratic equation admits up to two fixed points for  $p$ . We begin by examining the dynamics of  $p$  for the case perfectly balanced potentiation and depression ( $\delta = 0$ ) and a realistic shape of the STDP curve, and then consider the case of  $\delta \neq 0$ . We find that  $p$  potentiates for all cases except depression-dominated, balanced STDP, which can exhibit novel multistable dynamics.

Experimentally measured STDP rules in cortex often show  $f_+ > f_-$  and  $\tau_+ < \tau_-$  [55, 56], making potentiation windows sharper and higher-amplitude than depression windows. In this case, the STDP-weighted covariance from forward connections,  $S_F > 0$ , is greater in magnitude than those from backward connections,  $S_B < 0$  (Fig. 4), and hence  $S_F + p_0 S_B > 0$ . Furthermore, since the covariance from common input decays symmetrically around  $s = 0$  (Fig. 4C), we have that  $S_C > 0$ . Consequently, when  $\delta = 0$ , all terms in Eq. (11) are positive and  $p$  potentiates to  $p_0 W^{\max}$ .

For potentiation-dominated balanced STDP,  $+\delta\epsilon$ , again all terms in Eq. (11) are positive and  $p$  potentiates to  $p_0 W^{\max}$  (Fig. 6A). However, with depression-dominated balanced STDP ( $-\delta\epsilon$  in Eq. (11)),  $p$  has two fixed points, at:

$$p = \frac{-(S_F + p_0 S_B) \pm \sqrt{(S_F + p_0 S_B)^2 + 4\delta p_0 r^2 S_C}}{2S_C}. \quad (12)$$

Since  $(S_F + p_0 S_B) > 0$  and  $S_C > 0$  because of our assumptions on  $f_{\pm}$  and  $\tau_{\pm}$ , the term inside the square root is positive, and one fixed point is positive and the other negative. The positive fixed point is unstable and, if within  $[0, p_0 W^{\max}]$ , provides a separatrix between potentiation and depression of  $p$  (Fig. 6B). This separatrix arises from the competition between potentiation (due to forward connections and common input) and depression (due to backward connections and firing rates). Examination of Eq. (12) shows competing effects of increasing  $p_0$ : it moves both fixed points closer to 0 (the  $p_0 S_B$  terms outside and inside the square root), but also pushes the fixed points away from 0 due to firing rate- and common-input terms. The latter effect dominates for the positive fixed point. So the mean synaptic weight of sparsely connected networks have a propensity to potentiate, while more densely connected networks are more likely to depress (Fig. 6B).

In total, we see that a slight propensity for depression can impose bistability on the mean synaptic weight. In this case, a network with an initially strong mean synaptic weight  $p(0)$  can overcome depression and strengthen synaptic wiring, while a network with the same STDP rule and connection probability but with an initially weak mean synaptic weight will exhibit depression. In the next section we will show that similar separatrices exist in non-Erdős-Rényi networks with slightly depression-dominated STDP and govern motifs in such networks as well.

## Motif dynamics

We now consider networks that have structure at the level of motifs, so that different patterns of connectivity may be over- or under-represented compared to Erdős-Rényi networks. We begin by defining the two-synapse motif variables:

$$\begin{aligned} q^{\text{div}} &= \frac{1}{N^3} \sum_{i,j,k} \mathbf{w}_{ik} \mathbf{w}_{jk} - p^2, \\ q^{\text{con}} &= \frac{1}{N^3} \sum_{i,j,k} \mathbf{w}_{ik} \mathbf{w}_{ij} - p^2, \\ q^{\text{ch}} &= \frac{1}{N^3} \sum_{i,j,k} \mathbf{w}_{ij} \mathbf{w}_{jk} - p^2. \end{aligned} \quad (13)$$



The variables  $q^{\text{div}}$ ,  $q^{\text{con}}$  and  $q^{\text{ch}}$ , respectively, measure the strength of divergent, convergent, and chain motifs. For each variable, we subtract the expected value of the sum in a network with independent weights,  $p^2$ , so that the  $qs$  measure above- or below-chance levels of structure in the network. Since these variables depend on the strength of both synapses making up the motif, we will refer to them as *motif strengths*. Motif strengths are also related to neurons' (weighted) in- and out-degrees (the total strength of incoming or outgoing synapses for each neuron).  $q^{\text{div}}$  and  $q^{\text{con}}$  are proportional to the variance of neurons' in- and out-degrees.  $q^{\text{ch}}$ , on the other hand, is proportional to the covariance of neurons' in- and out-degrees. This can be seen by taking the definitions of these motifs, Eq. (13), and first summing over the indices  $i, j$ . This puts the sum in  $q^{\text{div}}$ , for example, in the form of a sum over neurons' squared out-degrees.)

We now wish to examine the joint dynamics of the mean synaptic weight  $p$  and these motif strengths. In order to calculate the dynamics of, for example,  $p$ , we insert the fast-slow STDP theory of Eq. (2) into the definition of  $p$ , as earlier. Similarly to Eq. (9), the dynamics of motifs  $q^{\text{div}}(t)$ ,  $q^{\text{con}}(t)$ , and  $q^{\text{ch}}(t)$  then depend on the network structure. This dependence of first- and second-order connectivity statistics on the network structure poses a challenge for the development of a closed theory for the dynamics of motifs. The main steps in developing such a theory are the two approximations we used to develop Eq. (9), as well as one more.

As above, our first approximation is to truncate the spike-train covariances at length one paths through the network. This removes the dependency of the dynamics on longer paths through the network. Nevertheless, after truncating  $\mathbf{C}(s)$ , the first- ( $p$ ) and second-order ( $q^{\text{div}}$ ,  $q^{\text{con}}$ ,  $q^{\text{ch}}$ ) motifs still depend on higher-order motifs (Eq. (8)). This is because of coupling between lower and higher-order moments of the connectivity matrix  $\mathbf{W}$  (see Eqs. (32)-(34)) and presents a significant complication.

In order to close the dynamics at one- and two-synapse motifs, our new approximation follows [16], and we rewrite higher-order motifs as combinations of individual synapses and two-synapse motifs (see Eqs. (39)-(40)). For the mean synaptic weight, for example, one third-order motif appears due to the common input term of the spike-train covariances (Eq. (8)). We break up this three-synapse motif into all possible combinations of two-synapse motifs and individual connections, estimating its strength as:

$$\frac{1}{N^3} \sum_{i,j,k} \mathbf{W}_{ij}^0 \mathbf{W}_{ik} \mathbf{W}_{jk} \approx \left( p_0 (q^{\text{div}} + p^2) + p (q_X^{\text{con}} + q_X^{\text{ch,B}}) \right). \quad (14)$$

This corresponds to assuming that there are no third- or higher-order correlations in the weight matrix beyond those due to second-order correlations; three- and more-synapse motifs are represented only as much as would be expected given the two-synapse motif strengths. This allows us to close the motif structure at two-synapse motifs. However, two new motifs appear in Eq. (14),  $q_X^{\text{con}}$  and  $q_X^{\text{ch,B}}$ . The  $x$  subscript denotes that these motifs are mixed between the weight and adjacency matrices, measuring the strength of individual connections, conditioned on their being part of a particular motif.  $q_X^{\text{con}}$  corresponds to the strength of connections conditioned on being part of a convergent motif and  $q_X^{\text{ch,B}}$  to the strength of connections conditioned on the postsynaptic neuron making another synapse in a chain (Eq. (26)). As in previous sections, the final approximation is to ignore the bounds on the synaptic weight in Eq. (1), so that our theory only captures the transient dynamics of  $\mathbf{W}(t)$ .

These approximations allow us (see Eqs. (29), (32), and (41)) to rewrite the dynamics of the mean synaptic weight  $p$  as:

$$\frac{dp}{dt} = p_0 r^2 S + \epsilon \left[ p S_F + (q_X^{\text{rec}} + p_0 p) S_B + \frac{1}{p_0} \left( p_0 (q^{\text{div}} + p^2) + p (q_X^{\text{con}} + q_X^{\text{ch,B}}) \right) S_C \right]. \quad (15)$$

The parameters  $S$ ,  $S_F$ ,  $S_B$  and  $S_C$  are defined as above. Note that we recover Eq. (9) when all  $q$ 's vanish (i.e an Erdős-Rényi network). When the network contains additional motifs ( $q \neq 0$ ), the dynamics of  $p$  contain new terms. In Eq. (15), the influence of forward connections through  $S_F$  is again proportional to the mean synaptic weight  $p$ . In contrast, the influence of backward connections  $S_B$  must interact with the new variable  $q_X^{\text{rec}}$ , which measures the mean strength of connections conditioned on their being part of a reciprocal loop (i.e the strength of a backwards connection, conditioned on the existence of the forward one). As described above (Eq. (14)), the covariance from common input  $S_C$  involves  $p$ , the divergent motif,  $q^{\text{div}}$ , as well as terms conditioned on weights being part of a convergent motif,  $q_X^{\text{con}}$ , or on the postsynaptic neuron making another synapse in a chain,  $q_X^{\text{ch,B}}$ . The definitions for the mixed motifs, the  $q_X$ s, are given in Eqs. (26). In total, the dynamics of mean synaptic weight cannot be written as a single closed equation, but also requires knowledge of how the second order motifs evolve.

Fortunately, using a similar approach, dynamical equations can be derived for each of the two-synapse motifs  $q^{\text{div}}$ ,  $q^{\text{cov}}$ , and  $q^{\text{ch}}$  (Eqs. (42)-(44)). However, to close the system we require dynamics for five mixed motifs,  $q_X^{\text{con}}$ ,  $q_X^{\text{div}}$ ,  $q_X^{\text{rec}}$ ,  $q_X^{\text{ch,A}}$ , and  $q_X^{\text{ch,B}}$  (Eqs. (45)-(49)). In total, this yields an autonomous 9-dimensional system of nonlinear differential equations describing the population averaged plasticity of second-order network structure. We have derived these equations in the absence of common external inputs to the neurons; the theory can easily be extended to this case by including external covariance in Eq. (7) (replacing  $\mathbf{C}^0$  with  $(\mathbf{C}^0 + \mathbf{C}^{\text{ext}})$ , where  $\mathbf{C}^{\text{ext}}$  is the covariance matrix of the inputs).

When the network structure,  $\mathbf{W}^0$  is approximately Erdős-Rényi, the motif frequencies  $q_0$  are  $\mathcal{O}(N^{-3/2}) = \mathcal{O}(\epsilon^{3/2})$ . If we further assume initial conditions for the motif strengths and the mixed motifs to be consistent with Erdős-Rényi statistics ( $q(0) \sim \mathcal{O}(\epsilon^{3/2})$  for all motifs), then we also have  $dq_X/dt \sim \mathcal{O}(\epsilon^{3/2})$  and  $dq_X/dt \sim \mathcal{O}(\epsilon^{3/2})$  for each motif, Eqs. (42)-(49). In this case we can neglect, to leading order, the dynamics of motifs entirely. Thus, the set of Erdős-Rényi networks  $\{p(t), q^{\text{div}} = q^{\text{con}} = q^{\text{ch}} = q_X^{\text{rec}} = q_X^{\text{con}} = q_X^{\text{div}} = q_X^{\text{ch,A}} = q_X^{\text{ch,B}} = 0\}$  forms an invariant set under the dynamics of the motif strengths; we examined the behavior on this invariant set in the above section (Fig. 5 and 6).

We refer to the mean field theory of Eqs. (41)-(49) as the *motif dynamics* for a recurrent network with STDP. This theory accurately predicts the transient dynamics of the first- and two-synapse motifs of the full stochastic spiking network (Fig. 7, compare red versus thin black curves), owing to significant drift compared to diffusion in the weight dynamics and these network-averaged motif strengths. The derivation and successful application of this reduced theory to a large spiking network is a central result of our study.

Our theory captures several nontrivial aspects of the evolution of network structure. First, while the STDP rule is in the depression-dominated regime ( $S < 0$  for the simulations in Fig. 7), the mean synaptic weight  $p$  nevertheless grows (Fig. 7A). Second, both divergent and convergent connections,  $q^{\text{div}}$  and  $q^{\text{con}}$ , grow above what is expected for a random (Erdős-Rényi) graph (Fig. 7B,C); however, at the expense of chain connections  $q^{\text{ch}}$  which decay (Fig. 7G). In the subsequent sections, we leverage the simplicity of our reduced theory to gain insight into how the STDP rule  $L(s)$  interacts with recurrent architecture to drive motif dynamics.

## Unbalanced STDP of two-synapse motifs

When the STDP rule is dominated by potentiation or depression so that  $S \sim \mathcal{O}(1) \gg \epsilon$ , then the  $\mathcal{O}(\epsilon)$  terms in Eqs. (42)-(49) are negligible. In this case each motif's plasticity is solely determined by the firing

rates, with spiking covariance plays no role. Here the motif dynamics are simply:

$$\begin{aligned}\frac{dp}{dt} &= p_0 r^2 S + \mathcal{O}(\epsilon) \\ \frac{dq^\alpha}{dt} &= 2r^2 S q_X^\alpha + \mathcal{O}(\epsilon) \\ \frac{dq_X^\alpha}{dt} &= r^2 S q_0^\alpha + \mathcal{O}(\epsilon)\end{aligned}\tag{16}$$

for  $\alpha = \text{div}, \text{con}, \text{orch}$  (and taking  $q_X^{\text{ch}} = (q_X^{\text{ch,A}} + q_X^{\text{ch,B}})/2$  in the second equation). The dynamics of  $p$  are the same here as for the Erdős-Rényi case above; we include it for completeness. Dropping order  $\epsilon$  terms gives the simple solutions:

$$\begin{aligned}p(t) &= p_0 r^2 S t + p(0) \\ q^\alpha(t) &= q^\alpha(0) + q_X^\alpha(0) r^2 S t + \frac{1}{2} q_0^\alpha (r^2 S)^2 t^2\end{aligned}\tag{17}$$

for  $\alpha = \text{div}, \text{con}, \text{ch}$  (Methods: Unbalanced STDP). As stated previously, with  $S \sim \mathcal{O}(1)$ , individual synapses uniformly potentiate or depress (Fig. 5). This is reflected in the linear decay or growth (for depression- or potentiation-dominated  $L(s)$ , respectively) of  $p$  with  $r^2$  and quadratic amplification of baseline motif frequencies for the two-synapse motif strengths.

## Balanced STDP of two-synapse motifs

Now, we turn our attention to how internally generated spiking covariance interacts with balanced STDP to control motifs (examining the dynamics of Eqs. (41)-(49)). As before, we consider STDP rules with sharper windows for potentiation than depression ( $\tau_+ < \tau_-$  and  $f_+ > f_-$ ). Each two-synapse motif can have a nullcline surface in the nine-dimensional motif space. These nullclines define a separatrix for the promotion or suppression of the corresponding motif, analogous to the case on the Erdős-Rényi invariant set (Fig. 7). We illustrate this by examining the dynamics in the  $(q^{\text{div}}, q^{\text{con}})$  plane. For STDP rules with a balance tilted towards depression ( $-\delta\epsilon$ ), the nullclines provided thresholds for the promotion or suppression of divergent or convergent motifs (Fig. 8A, blue lines). The flow in this slice of the motif space predicted the motif dynamics well (Fig. 8A, compare individual realizations of the full spiking network – thin black lines – to the flow defined by the vector field of the reduced motif system).

For STDP rules with the balance tilted towards potentiation ( $+\delta\epsilon$ ), on the other hand, the nullclines were at negative motif strengths (Fig. 8B). Can the motif strengths achieve negative values? As stated previously,  $q^{\text{con}}$  and  $q^{\text{div}}$  are proportional to the variances of neurons' in and out degrees, respectively. So, like the mean synaptic weight,  $q^{\text{div}}, q^{\text{con}} \geq 0$ , and these motifs always potentiated for  $+\delta\epsilon$  STDP rules (Fig. 8B).

Chain motifs, in contrast, correspond to the covariance of neurons' weighted in- and out-degrees and so can achieve negative values. Indeed, the strength of chains can depress below zero even while the mean synaptic weight and other motifs potentiate (Fig. 5A,G). Examining how  $q^{\text{ch}}, q^{\text{div}}$  and  $q^{\text{con}}$  coevolve allowed us to see how in- and out-hubs developed in the network. With the  $+\delta\epsilon$  STDP rule,  $q^{\text{ch}}$  increased along with  $q^{\text{con}}$  and  $q^{\text{div}}$  (Figs. 8B, 9C,D). So, individual neurons tended to become both in- and out-hubs. With the  $-\delta\epsilon$  STDP rule, however,  $q^{\text{ch}}$  could decrease while  $q^{\text{div}}$  and  $q^{\text{con}}$  increased (Fig. 5). In this case, neurons tended to become in- or out-hubs, but not both.

Many studies have examined how STDP affects either feedforward or recurrent structure in neuronal networks, commonly showing that STDP promotes feedforward structure at the expense of recurrent loops [36, 57, 58]. This is consistent with the intuition gained from isolated pairs of neurons, where STDP can induce competition between reciprocal synapses and eliminate disynaptic loops (Supp. Fig. 1; [24]). Our theory provides a new way to examine how STDP regulates feedforward vs recurrent motifs by examining the dynamics of  $q^{\text{ch}}$ . This variable includes both recurrent loops ( $q^{\text{rec}}$ ) and feedforward chains ( $q^{\text{ff}}$ ). In order to understand the contribution of each of these to overall potentiation or depression of chains, we split the motif strength for chains into contributions from recurrent loops and feedforward chains, rewriting  $q^{\text{ch}}$  as:

$$q^{\text{ch}} = \underbrace{\frac{1}{N^3} \sum_{i,j,k} \delta_{ik} \mathbf{W}_{ij} \mathbf{W}_{jk}}_{q^{\text{rec}}} + \underbrace{\frac{1}{N^3} \sum_{i,j,k} (1 - \delta_{ik}) \mathbf{W}_{ij} \mathbf{W}_{jk}}_{q^{\text{ff}}} - p^2. \quad (18)$$

Similar to the case of other two-synapse motifs, the leading order dynamics of the recurrent motif are:

$$\frac{1}{2\epsilon} \frac{dq^{\text{rec}}}{dt} = r^2 S p_0 (q_{\text{X}2}^{\text{rec}} + p p_0) + S_F q^{\text{rec}} + S_B q_{\text{X}2}^{\text{rec}}. \quad (19)$$

We obtain the dynamics of the feedforward motif by subtracting  $dq^{\text{rec}}/dt$  from  $dq^{\text{ch}}/dt$  (Eq. (54)). In Eq. (18) we subtract  $p^2$  from  $q^{\text{ff}}$  because  $q^{\text{ff}}$  is the dominant contributor to  $q^{\text{ch}}$ . This restricts  $q^{\text{rec}}$  to being non-negative. The new auxiliary variable  $q_{\text{X}2}^{\text{rec}}$  is proportional to the conditional second moment of weights that are part of loops (Eq. (51)), and evolves according to Eq. (53). The replacement of  $q^{\text{ch}}$  by these variables expands the motif space to 11 dimensions.

We investigated the joint dynamics of feedforward chains and recurrent loops similarly to the other motifs, examining the  $(q^{\text{ff}}, q^{\text{rec}})$  plane. The  $q^{\text{ff}}$  and  $q^{\text{rec}}$  nullclines divided this plane into regions where each motif potentiated or depressed. The shape of the STDP rule and the initial values of the other motif strengths determined the location of these nullclines. For the  $+\delta\epsilon$  STDP rule, the  $q^{\text{rec}}$  nullcline was just below  $q^{\text{rec}} = 0$  (Fig. 9A, blue horizontal line). Since  $q^{\text{rec}} \geq 0$ , this forced  $q^{\text{rec}}$  to potentiate. The feedforward motif, in contrast, could potentiate or depress above chance levels. In our spiking simulations, the initial conditions put  $q^{\text{ff}}$  in the region of depression, so that feedforward structure depressed even while all other motifs were growing (Fig. 9A, right panels).

These dynamics were the opposite of what would be expected from examining isolated pairs of neurons. With both the  $+\delta\epsilon$  and  $-\delta\epsilon$  balanced STDP rules, isolated pairs of neurons showed splitting of synaptic weights to eliminate the recurrent loop (Supp. Fig. 1). Thus, with the  $+\delta\epsilon$  STDP rule, the intuition gained from pairs of neurons did not predict the plasticity of feedforward and recurrent motifs.

The locations of the  $q^{\text{ff}}$  and  $q^{\text{rec}}$  nullclines were sensitive to the values of the other motif variables. Since the mean synaptic weight and  $q^{\text{div}}$  and  $q^{\text{con}}$  exhibited bistability under the  $-\delta\epsilon$  STDP rule, we examined the  $(q^{\text{ff}}, q^{\text{rec}})$  slice through motif space when the other motifs were potentiating (Fig. 9B, right panels) or depressing (Fig. 9C, right panels). In both cases, the  $q^{\text{rec}}$  nullcline was above 0 so that the recurrent motif could either potentiate or depress, depending on its initial strength (Fig. 9B,C blue horizontal lines). Similarly, the feedforward motif could either potentiate or depress.

In spiking simulations with  $-\delta\epsilon$  STDP where  $p$  and the other motifs potentiated (Fig. 9B, right), the initial conditions put  $(q^{\text{ff}}, q^{\text{rec}})$  in the region of phase space where they both depressed (Fig. 9B, left). In spiking simulations with  $-\delta\epsilon$  STDP where  $p$  and other motifs depressed (Fig. 9C, right), the initial conditions put  $(q^{\text{ff}}, q^{\text{rec}})$  in the region where  $q^{\text{ff}}$  potentiated and  $q^{\text{rec}}$  depressed. This region corresponded

to what would be expected from examining pairs of neurons (Supp. Fig. 1): the loss of disynaptic loops and promotion of feedforward structure. So with the  $-\delta\epsilon$  STDP rule, the region of phase space where the pair-based intuition was accurate at the network level was accessible. In most of the motif space, however, interactions between triplets of neurons played a strong role so that the theory developed here was necessary to predict the STDP of motif structure.

## Discussion

We have developed a theory for spike timing-dependent plasticity in weakly coupled recurrent networks of exponential integrate-and-fire neurons. We used this framework to derive a low-dimensional dynamical system capturing the plasticity of two-synapse motifs. The resulting system naturally classifies STDP rules into two categories: 1) rules with an imbalance between potentiation and depression whose dynamics are dominated by the firing rates of neurons in the network, and 2) rules with balanced potentiation and depression in which different sources of spiking covariance interact with the STDP rule to determine network structure. In the latter case, any mechanism controlling spiking covariance in the network may affect how the network structure evolves. Thus, spike initiation dynamics [59–62], spike-frequency adaptation [63, 64], synaptic inhibition [65–67] and passive membrane properties [68] could all, in addition to controlling firing rates, drive motif dynamics.

### STDP in recurrent networks

A recent suite of studies derived a theory for how STDP shapes the full structure of networks of neurons whose spike trains are Poisson processes [35, 43–46, 54]. The initial approach is similar to ours with a linear approximation to estimate spiking covariance (see Eq. (3)-(4)). However, these studies mostly focused on the effects of external input, considering how feedforward inputs entrain structure in recurrent synapses [35, 44, 54]. The only source of spiking covariance was inherited from external sources (meaning  $\mathbf{C}_0(s)$  has off-diagonal structure), and subsequently filtered by the network to produce spiking covariance. Two previous studies by the same authors also examined STDP in networks without external stimuli [43, 45]; however, these took a large system size limit ( $N \rightarrow \infty$ ) and neglected the contribution of spiking covariance to STDP, focusing on the firing rate dependence due to an unbalanced learning rule.

In contrast, we consider the case where the intrinsic variability of neurons’ spike trains is the only source of spiking covariance, necessitating a finite sized network ( $\epsilon = 1/(Np_0) > 0$ ). There is little difference between our results and those of past studies [43, 45] when the learning rule is unbalanced. However, if there is a balance between potentiation and depression, our theory shows how internally generated spiking covariances play a strong role in STDP. Furthermore, our use of integrate-and-fire models allows our theory to predict the evolution of network structure without fixing the statistics of individual or joint spiking activity.

### Stability of learned network structures

Early studies of long-term plasticity, which gave rise to the phenomenological plasticity model we used, focused on the relative timing of action potentials. More recent experiments have shown that neurons’ firing rates and the postsynaptic membrane voltage and spike patterns all affect the shape of measured STDP curves [56, 69–72]. More complicated models of long-term plasticity, based on spike-triplet- or voltage-dependent STDP [73, 74] or on calcium thresholds for the induction of depression and potentiation [75–77], can replicate many of these complexities. The observation that firing rates undergo large fluctuations over slow timescales [78–82] suggests that *in vivo* STDP may transition between unbalanced potentiation- and depression-dominated regimes. While long-term plasticity can be strongly

affected by pre- and postsynaptic firing rates, connectivity motifs and spiking covariance could determine the direction of plasticity during transitions between potentiation- and depression-dominated regimes. While our paper provides an initial framework to study how STDP shapes structure in recurrent networks, a more realistic learning rule than that used here (Eq. (1)) will be needed to address these issues.

The additive, Hebbian STDP model we used here gives rise to splitting of synaptic weights: individual weights potentiate to some upper bound, or depress to a lower bound. This produces a bimodal distribution of synaptic weights, while experimentally observed weight distributions tend to be unimodal and long-tailed [3, 4, 83, 84]. Modifications of this model, such as introducing axonal or dendritic delays or weight-dependence of plasticity, can yield weight distributions more closely resembling those observed in neural tissue [30–32, 85, 86]. Depending on the modification made (delays vs weight-dependence), either the same or similar theories for motif plasticity can be derived using the methods presented in our study. Strong weight dependence, however, forces every weight to the same value so that the baseline motif frequencies completely determine the structure of the weight matrix (Supplemental Information: Multiplicative STDP). The dynamics of motifs under more realistic models of synaptic plasticity remain to be studied.

A major feature of STDP is that it can potentiate temporally correlated inputs [33]. Since synchronous inputs are effective at driving postsynaptic spiking, this can give rise to pathological activity in recurrent networks [39]. Synaptic depression driven by postsynaptic spikes, independent of presynaptic activity, can stabilize postsynaptic firing rates during STDP [29, 35]. Such additional rate-dependent terms of the plasticity rule can also stabilize the full weight matrix [45] and thus give rise to stable motif configurations. Recent work has focused on the necessity of homeostatic mechanisms, including synaptic scaling [87] or inhibitory plasticity, in stabilizing both the activity and structure of neural networks [36, 41, 88–91]. Since balanced STDP can give rise to bistability of mean synaptic weights in a network (Fig. 7B), it could also provide a mechanism for assembly formation (selected weights potentiate, while other weights depress). Mechanisms of metaplasticity [92], operating on a similar timescale to STDP, could give rise to such a balance. This suggests a novel role for metaplasticity in controlling not only single-neuron excitability but also the self-organization of microcircuits in recurrent networks.

## Plasticity of motifs

Early studies on STDP focused on isolated pairs of reciprocally connected neurons, showing that the type of STDP we study tends to induce competition between reciprocal synapses (Fig. 1B,C; [24]). Since then, many simulation studies have investigated how STDP affects the structure and activity of recurrent networks [38, 41, 93–95], commonly examining the emergence of highly connected clusters. Reduced theories exposing how STDP shapes network-level structure have, however, been difficult to obtain. Most have examined the average synaptic weight in a network [96, 97], focusing on the relationship between network-averaged firing rates and mean synaptic weights ( $p$ ) but neglecting spiking covariance. Mean-field theories are accurate for fully homogenous networks; however, if all neurons have the same weighted in- and out-degrees, there is no plasticity of two-synapse motifs (Supplemental Information: Motif plasticity in homogenous networks). So plasticity of higher-order network structure depends on inhomogeneities in neurons’ inputs and outputs.

The few reduced theories examining STDP of higher-order network structure have focused on the question of how STDP controls feedforward chains versus recurrent loops. One study compared the mean strengths of feedforward versus recurrent inputs in a network receiving synchronous stimulation [58], but did so for a neuron that made no feedback connections to the network – effectively only taking into account the first term of Eq. (7). Another study examined the strength of loops in a network of linear

excitatory neurons, showing that STDP tends to reduce the total number of loops (of all lengths) in a network [57]. Our theory is restricted to two-synapse loops; while we have shown that these can potentiate (as in Fig. 9C), [57] predicts that longer loops would meanwhile be weakened. Whether this is the case with balanced STDP driven by more realistic neuron models remains to be seen.

There is a growing body of evidence that cortical networks exhibit fine-scale structure [2–5]. Experimental studies have shown that such microcircuits depend on sensory experience [98, 99]. Our work provides an important advance towards explicitly linking the plasticity rules that control individual synapses and the emergent microcircuits of cortical networks. We have shown that synaptic plasticity based only on temporally precise spike-train covariance can give rise to a diversity and, under certain conditions, multistability of motif configurations. Motifs can have a strong influence on pairwise and population-level activity [8–18], suggesting that precise spike timing may play a role in how networks reorganize patterns of connectivity in order to learn computations.

## Methods

### Neuron and network model

We model a network of  $N$  neurons. The membrane dynamics of individual neurons obey the exponential integrate-and-fire (EIF) model [50], one of a class of models well-known to capture the spike initiation dynamics and statistics of cortical neurons [51, 52]. Specifically, the membrane voltage of neuron  $i$  evolves according to:

$$C \frac{dV_i}{dt} = g_L (V_L - V_i) + g_L \Delta \exp\left(\frac{V_i - V_T}{\Delta}\right) + I_i(t) + \sum_{j=1}^N \mathbf{W}_{ij} (\mathbf{J}_{ij} * y_{j\cdot}). \quad (20)$$

The first term on the right-hand side is the leak current, with conductance  $g_L$  and reversal potential  $V_L$ . The next term describes a phenomenological action potential with an initiation threshold  $V_T$  and steepness  $\Delta$ : when the voltage reaches  $V_T$ , it diverges; this divergence marks an action potential. For numerical simulations, action potentials are thresholded at  $V(t) = V_{th}$ , reset to a reset potential  $V_{re}$  and held there for an absolute refractory period  $\tau_{ref}$ .

Input from external sources not included in the model network is contained in  $I_i(t)$ . We model this as a Gaussian white noise process:  $I_i(t) = \mu + g_L \sigma D \xi_i(t)$ . The mean of the the external input current is  $\mu$ . The parameter  $\sigma$  controls the strength of the noise and  $D = \sqrt{\frac{2C}{g_L}}$  scales the noise amplitude to be independent of the passive membrane time constant. With this scaling, the infinitesimal variance of the passive membrane voltage is  $(g_L \sigma D)^2$ .

The last term of Eq. (20) models synaptic interactions in the network. The  $N \times N$  matrix  $\mathbf{W}$  contains the amplitudes of each synapse's postsynaptic currents. It is a weighted version of the binary adjacency matrix  $\mathbf{W}^0$ , where  $\mathbf{W}_{ij}^0 = 1(0)$  indicates the presence (absence) of a synapse from neuron  $j$  onto neuron  $i$ . If a synapse  $ij$  is present then  $\mathbf{W}_{ij}$  denotes its strength. Due to synaptic plasticity,  $\mathbf{W}$  is dynamic; it changes in time as individual synapses potentiate or depress. The spike train from neuron  $j$  is the point process  $y_j(t) = \sum_k \delta(t - t_j^k)$ , where  $t_j^k$  denotes the  $k^{\text{th}}$  spike time from neuron  $j$ . The  $N \times N$  matrix  $\mathbf{J}(t)$  defines the shape of the postsynaptic currents. In this study, we use exponential synapses:  $\mathbf{J}_{ij}(t - t_j^k) = \mathcal{H}(t - t_j^k) \exp\left(-\frac{t - t_j^k}{\tau_s}\right)$ , where  $\mathcal{H}(t)$  is the Heaviside step function. Our theory is not exclusive to the EIF model or to the simple synaptic kernels we used; similar methods can be used with

any integrate-and-fire model and arbitrary synaptic kernels. Model parameters are contained in Table 1 (unless specified otherwise in the text).

Unless otherwise stated we take the adjacency matrix  $\mathbf{W}_0$  to have Erdős-Rényi statistics with connection probability  $p_0 = 0.15$ .

## Learning dynamics

We now derive Eq. (2), summarizing a key result of [33]. Changes in a synaptic weight  $\mathbf{W}_{ij}$  are governed by the learning rule  $L(s)$ , Eq. (1). We begin by considering the total change in synaptic weight during an interval of length  $T$  ms:

$$\Delta \mathbf{W}_{ij} = \mathbf{W}_{ij}^0 \int_t^{t+T} \int_t^{t+T} L(t'' - t') y_j(t'') y_i(t') dt'' dt' \quad (21)$$

where multiplying by the corresponding element of the adjacency matrix ensures that nonexistent synapses do not potentiate into existence. Consider the trial-averaged rate of change:

$$\frac{\langle \Delta \mathbf{W}_{ij} \rangle}{T} = \mathbf{W}_{ij}^0 \frac{1}{T} \int_t^{t+T} \int_{t-t'}^{t+T-t'} L(s) \langle y_j(t' + s) y_i(t') \rangle ds dt' \quad (22)$$

where  $s = t'' - t'$  and  $\langle \cdot \rangle$  denotes the trial average. We first note that this contains the definition of the trial-averaged spike train cross-covariance:

$$\mathbf{C}_{ij}(s) = \frac{1}{T} \int_t^{t+T} \langle y_j(t' + s) y_i(t') \rangle dt' - r_i r_j \quad (23)$$

where  $r_i$  is the time-averaged firing rate of neuron  $i$  and subtracting off the product of the rates corrects for chance spike coincidences. Inserting this definition into Eq. (22) yields:

$$\frac{\langle \Delta \mathbf{W}_{ij} \rangle}{T} = \mathbf{W}_{ij}^0 \int_{t-t'}^{t+T-t'} L(s) (r_i r_j + \mathbf{C}_{ij}(s)) ds \quad (24)$$

We then take the amplitude of individual changes in the synaptic weights to be small:  $f_+, f_- \ll W^{\max}$ , where  $\tau_{\pm}$  define the temporal shape of the STDP rule (see Eq. (1)). In this case, changes in the weights occur on a slower timescale than the width of the learning rule. Taking  $T \gg \max(\tau_+, \tau_-)$  allows us to extend the limits of integration in Eq. (24) to  $\pm\infty$ , which gives Eq. (2). Note that in the results we have dropped the angle brackets for convenience. This can also be justified by the fact that the plasticity is self-averaging, since  $\Delta \mathbf{W}_{ij}$  depends on the integrated changes over the period  $T$ .

## Spiking statistics

In order to calculate  $d\mathbf{W}_{ij}/dt$ , we need to know the firing rates  $r_i, r_j$  and spike train cross-covariance  $\mathbf{C}_{ij}(s)$  (Eq. (2)). We take the weights to be constant on the fast timescale of  $s$ , so that the firing rates and spike train cross-covariances are stationary on that timescale. We solve for the baseline firing rates in the network via the self-consistency relationship

$$r_i = r_i(\mu_i^{\text{eff}}, \sigma), \text{ where} \\ \mu_i^{\text{eff}} = \mu + \sum_j \left( \int_{-\infty}^{\infty} \mathbf{J}_{ij}(t) dt \right) \mathbf{W}_{ij} r_j$$

for  $i = 1, \dots, N$ . This gives the equilibrium state of each neuron's activity. In order to calculate the spike train cross-covariances, we must consider temporal fluctuations around the baseline firing rates.



With sufficiently weak synapses compared to the background input, we can linearize each neuron's activity around the baseline state. Rather than linearizing each neuron's firing rate around  $r_i$ , we follow [15, 47, 48] and linearize each neuron's spike train around a realization of background activity, the uncoupled spike train  $\mathbf{y}_i^0$  (Eq. (3)). The perturbation around the background activity is given by each neuron's linear response function,  $\mathbf{A}_i(t)$ , which measures the amplitude of firing rate fluctuations in response to perturbations of each neuron's input around the baseline  $\mu_i^{\text{eff}}$ . We calculate  $\mathbf{A}(t)$  using standard methods based on Fokker-Planck theory for the distribution of a neuron's membrane potential [100, 101].

This yields Eq. (3), approximating a realization of each neuron's spike train as a mixed point and continuous process. Spike trains are defined, however, as pure point processes. Fortunately, Eq. (2) shows that we do not need a prediction of individual spike train realizations, but rather of the trial-averaged spiking statistics. We can solve Eq. (3) for the spike trains in the frequency domain as:

$$\mathbf{y}(\omega) = (\mathbf{I} - (\mathbf{W} \cdot \mathbf{K}(\omega)))^{-1} \mathbf{y}^0(\omega)$$

where as in the Results,  $\mathbf{K}(\omega)$  is an interaction matrix defined by  $\mathbf{K}_{ij}(\omega) = \mathbf{A}_i(\omega) \mathbf{J}_{ij}(\omega)$  and  $\cdot$  denotes the element-wise product. Averaging this expression over realizations of the background spike trains yields a linear equation for the instantaneous firing rates. Averaging the spike trains  $\mathbf{y}$  against each other yields the full cross-covariance matrix, Eq. (4). It depends on the coupling strengths  $\mathbf{W}$ , the synaptic filters  $\mathbf{J}_{ij}$  and neurons' linear response functions  $\mathbf{A}$ , and the covariance of the baseline spike trains,  $\mathbf{C}^0$ .

We can calculate the baseline covariance in the frequency domain,  $\mathbf{C}^0(\omega) = \langle \mathbf{y}^0 \mathbf{y}^{0*} \rangle$ , by first noting that it is a diagonal matrix containing each neuron's spike train power spectrum. We calculate these using the renewal relationship between the spike train power spectrum  $\mathbf{C}^0(\omega)$  and the first passage time density [102]; the first passage time density for nonlinear integrate and fire models can be calculated using similar methods as for the linear response functions [101].

## Self-consistent theory for network plasticity

We solve the system Eqs. (2),(4) for the evolution of each synaptic weight with the Euler method with a time step of 100 seconds. A package of code for solving the self-consistent theory and running the spiking simulations, in MATLAB and C, is available at <http://sites.google.com/site/gabrielkochocker/code>. Additional code is available on request.

## Derivation of motif dynamics

The baseline structure of the network is defined by the adjacency matrix  $\mathbf{W}^0$ . The frequencies of different motifs are:

$$\begin{aligned} p_0 &= \frac{1}{N^2} \sum_{i,j} \mathbf{w}_{ij}^0, \\ q_0^{\text{div}} &= \frac{1}{N^3} \sum_{i,j,k} \mathbf{w}_{ik}^0 \mathbf{w}_{jk}^0 - p_0^2, \\ q_0^{\text{con}} &= \frac{1}{N^3} \sum_{i,j,k} \mathbf{w}_{ik}^0 \mathbf{w}_{ij}^0 - p_0^2, \\ q_0^{\text{ch}} &= \frac{1}{N^3} \sum_{i,j,k} \mathbf{w}_{ij}^0 \mathbf{w}_{jk}^0 - p_0^2, \\ q_0^{\text{rec}} &= \frac{1}{N^2} \sum_{i,j} \mathbf{w}_{ij}^0 \mathbf{w}_{ji}^0 - p_0^2. \end{aligned} \tag{25}$$

Each of the  $q_0$  parameters refers to a different two-synapse motif. In divergent motifs ( $q_0^{\text{div}}$ ), one neuron  $k$  projects to two others,  $i$  and  $j$ . In convergent motifs ( $q_0^{\text{con}}$ ), two neurons  $k$  and  $j$  project to a third,  $i$ . In chain motifs ( $q_0^{\text{ch}}$ ), neuron  $k$  projects to neuron  $j$ , which projects to neuron  $i$ . Finally, in recurrent motifs ( $q_0^{\text{rec}}$ ) two neurons connect reciprocally. In each of these equations, we subtract off  $p_0^2$  to correct for the baseline frequencies expected in Erdős-Rényi random networks. So, these parameters measure above-chance levels of motifs in the adjacency matrix  $\mathbf{W}^0$ .

We extend this motif definition to a weighted version, given by Eqs. (13). Since our linear response theory for synaptic plasticity requires weak synapses, here we explicitly scale by the mean in-degree  $\epsilon = \frac{1}{Np_0}$ :

$$\begin{aligned}
\epsilon p &= \frac{1}{N^2} \sum_{i,j} \mathbf{w}_{ij}, \\
\epsilon^2 q^{\text{div}} &= \frac{1}{N^3} \sum_{i,j,k} \mathbf{w}_{ik} \mathbf{w}_{jk} - \epsilon^2 p^2, \\
\epsilon^2 q^{\text{con}} &= \frac{1}{N^3} \sum_{i,j,k} \mathbf{w}_{ik} \mathbf{w}_{ij} - \epsilon^2 p^2, \\
\epsilon^2 q^{\text{ch}} &= \frac{1}{N^3} \sum_{i,j,k} \mathbf{w}_{ij} \mathbf{w}_{jk} - \epsilon^2 p^2, \\
\epsilon q_X^{\text{rec}} &= \frac{1}{N^2} \sum_{i,j} \mathbf{w}_{ij} \mathbf{w}_{ji}^0 - \epsilon p p_0, \\
\epsilon q_X^{\text{div}} &= \frac{1}{N^3} \sum_{i,j,k} \mathbf{w}_{ik} \mathbf{w}_{jk}^0 - \epsilon p p_0, \\
\epsilon q_X^{\text{con}} &= \frac{1}{N^3} \sum_{i,j,k} \mathbf{w}_{ik} \mathbf{w}_{ij}^0 - \epsilon p p_0, \\
\epsilon q_X^{\text{ch,A}} &= \frac{1}{N^3} \sum_{i,j,k} \mathbf{w}_{ij} \mathbf{w}_{jk}^0 - \epsilon p p_0, \\
\epsilon q_X^{\text{ch,B}} &= \frac{1}{N^3} \sum_{i,j,k} \mathbf{w}_{ij}^0 \mathbf{w}_{jk} - \epsilon p p_0
\end{aligned} \tag{26}$$

Here we have defined the two-synapse motifs, as well as five auxiliary variables,  $\{q_X\}$ . These mixed motifs, defined by products of the weight and adjacency matrices, measure the strength of synapses *conditioned* on their being part of a motif. The motifs  $\{q\}$ , on the other hand, measure the total strength of the motifs. While the variables  $\{q_X\}$  are not of direct interest, we will see that they are required in order to close the system of equations. In comparison to the motif *frequencies*  $\{q_0\}$ , which measure motif frequencies in comparison to an independently *connected* network, the motif *strengths* are defined relative to an independently *weighted* network.

We also scale the amplitude of individual synaptic changes,  $L(s)$ , by  $\epsilon$ . We now go through the derivation of  $dp/dt$ ,  $dq^{\text{div}}/dt$  and  $dq_X^{\text{div}}/dt$  as examples; the other six variables follow the same steps. First, note that the spike train cross-covariance matrix of the network, Eq. (4), can be expanded in the Fourier domain around the baseline covariance  $\mathbf{C}^0(\omega)$ :

$$\mathbf{C}(\omega) = \left( \sum_{i=0}^{\infty} (\mathbf{W} \cdot \mathbf{K})^i \right) \mathbf{C}^0(\omega) \left( \sum_{j=0}^{\infty} ((\mathbf{W} \cdot \mathbf{K})^*)^j \right) \tag{27}$$

where the interaction matrix  $\mathbf{W} \cdot \mathbf{K}$  is the element-wise product of the weight matrix  $\mathbf{W}$  and the matrix of filters,  $\mathbf{K}$ . Powers of  $\mathbf{W} \cdot \mathbf{K}$  represent lengths of paths through the network. Only taking into account up to length one paths yields (for  $i \neq j$ ):

$$\mathbf{C}_{ij}(s) \approx \underbrace{(\mathbf{W}_{ij}\mathbf{K}_{ij} * \mathbf{C}_{jj}^0)(s)}_{\text{forwardconnection}} + \underbrace{(\mathbf{C}_{ii}^0 * \mathbf{W}_{ji}\mathbf{K}_{ji}^-)(s)}_{\text{backwardconnection}} + \underbrace{\sum_k (\mathbf{W}_{ik}\mathbf{K}_{ik} * \mathbf{C}_{kk}^0 * \mathbf{W}_{jk}\mathbf{K}_{jk}^-)(s)}_{\text{commoninputs}}. \quad (28)$$

where we have inverse Fourier transformed for convenience in the following derivation and  $\mathbf{K}^-(t) = \mathbf{K}(-t)$ .

Differentiating each motif with respect to time, using the fast-slow STDP theory Eq. (2) and inserting the first-order truncation of the cross-covariance functions, Eq. (7), yields:

$$\epsilon \frac{dp}{dt} = \frac{1}{N^2} \sum_{i,j} \mathbf{W}_{ij}^0 \int_{-\infty}^{\infty} \epsilon L(s) \left( r_i r_j + \delta_{ij} \mathbf{C}_{ij}^0(s) + (\mathbf{W}_{ij}\mathbf{K}_{ij} * \mathbf{C}_{jj}^0)(s) \right. \\ \left. + (\mathbf{C}_{ii}^0 * \mathbf{W}_{ji}\mathbf{K}_{ji}^-)(s) + \sum_k (\mathbf{W}_{ik}\mathbf{K}_{ik} * \mathbf{C}_{kk}^0 * \mathbf{W}_{jk}\mathbf{K}_{jk}^-)(s) \right) ds \quad (29)$$

$$\epsilon^2 \frac{dq^{\text{div}}}{dt} = \frac{2}{N^3} \sum_{i,j,k} \left[ \mathbf{W}_{ik}\mathbf{W}_{jk}^0 \int_{-\infty}^{\infty} \epsilon L(s) \left( r_j r_k + \delta_{jk} \mathbf{C}_{jk}^0(s) + (\mathbf{W}_{jk}\mathbf{K}_{jk} * \mathbf{C}_{kk}^0)(s) \right. \right. \\ \left. \left. + (\mathbf{C}_{jj}^0 * \mathbf{W}_{kj}\mathbf{K}_{kj}^-)(s) + \sum_l (\mathbf{W}_{jl}\mathbf{K}_{jl} * \mathbf{C}_{ll}^0 * \mathbf{W}_{kl}\mathbf{K}_{kl}^-)(s) \right) ds \right] - 2\epsilon^2 p \frac{dp}{dt} \quad (30)$$

$$\epsilon \frac{dq_X^{\text{div}}}{dt} = \frac{1}{N^3} \sum_{i,j,k} \mathbf{W}_{jk}^0 \mathbf{W}_{ik}^0 \int_{-\infty}^{\infty} \epsilon L(s) \left( r_i r_k + \delta_{ik} \mathbf{C}_{ik}^0(s) + (\mathbf{W}_{ik}\mathbf{K}_{ik} * \mathbf{C}_{kk}^0)(s) \right. \\ \left. + (\mathbf{C}_{ii}^0 * \mathbf{W}_{ki}\mathbf{K}_{ki}^-)(s) + \sum_l (\mathbf{W}_{il}\mathbf{K}_{il} * \mathbf{C}_{ll}^0 * \mathbf{W}_{kl}\mathbf{K}_{kl}^-)(s) \right) ds \Big] - \epsilon p_0 \frac{dp}{dt} \quad (31)$$

We now assume that all neurons have the same firing rates, spike train autocovariances and linear response functions:  $\forall i, r_i \equiv r$ ,  $\mathbf{C}_{ii}^0 \equiv \mathbf{C}^0$  and  $\mathbf{A}_i \equiv \mathbf{A}$ . Since we model all postsynaptic currents with the same shape, this makes the matrix  $\mathbf{K}$  a constant matrix; we replace its elements with the scalar  $K$ . Also neglecting the weight bounds in  $L(s)$  allows us to write:

$$\frac{dp}{dt} = r^2 S \frac{1}{N^2} \sum_{i,j} \mathbf{W}_{ij}^0 + S_F \frac{1}{N^2} \sum_{i,j} \mathbf{W}_{ij}^0 \mathbf{W}_{ij} + S_B \frac{1}{N^2} \sum_{i,j} \mathbf{W}_{ij}^0 \mathbf{W}_{ji} + S_C \frac{1}{N^2} \sum_{i,j,k} \mathbf{W}_{ij}^0 \mathbf{W}_{ik} \mathbf{W}_{jk} \quad (32)$$

$$\epsilon \frac{dq^{\text{div}}}{dt} = r^2 S \frac{2}{N^3} \sum_{i,j,k} \mathbf{W}_{ik} \mathbf{W}_{jk}^0 + S_F \frac{2}{N^3} \sum_{i,j,k} \mathbf{W}_{ik} \mathbf{W}_{jk}^0 \mathbf{W}_{jk} \\ + S_B \frac{2}{N^3} \sum_{i,j,k} \mathbf{W}_{ik} \mathbf{W}_{jk}^0 \mathbf{W}_{kj} + S_C \frac{2}{N^3} \sum_{i,j,k,l} \mathbf{W}_{ik} \mathbf{W}_{jk}^0 \mathbf{W}_{jl} \mathbf{W}_{kl} - 2\epsilon p \frac{dp}{dt} \quad (33)$$

$$\frac{dq_X^{\text{div}}}{dt} = r^2 S \frac{1}{N^3} \sum_{i,j,k} \mathbf{W}_{jk}^0 \mathbf{W}_{ik}^0 + S_F \frac{1}{N^3} \sum_{i,j,k} \mathbf{W}_{jk}^0 \mathbf{W}_{ik}^0 \mathbf{W}_{ik} \\ + S_B \frac{1}{N^3} \sum_{i,j,k} \mathbf{W}_{jk}^0 \mathbf{W}_{ik}^0 \mathbf{W}_{ki} + S_C \frac{1}{N^3} \sum_{i,j,k,l} \mathbf{W}_{jk}^0 \mathbf{W}_{ik}^0 \mathbf{W}_{il} \mathbf{W}_{kl} - p_0 \frac{dp}{dt} \quad (34)$$

where we have cancelled off an  $\epsilon$  from the left and right-hand sides. We have absorbed the integrals over the STDP rule and the spiking covariances into  $r^2 S$ ,  $S_F$ ,  $S_B$  and  $S_C$ . These correspond, respectively, to

the total STDP-weighted spiking covariances from chance coincidence, forward connections, backward connections, and common input:

$$S = \int_{-\infty}^{\infty} L(s) ds \quad (35)$$

$$S_F = \int_{-\infty}^{\infty} L(s) (K(t) * C^0(s)) ds \quad (36)$$

$$S_B = \int_{-\infty}^{\infty} L(s) (C^0(s) * K^-(t)) ds \quad (37)$$

$$S_C = \int_{-\infty}^{\infty} L(s) (K(t) * C^0(s) * K^-(t)) ds \quad (38)$$

These parameters depend on the spike train auto-covariance  $C^0(s)$  and linear response functions  $A(t)$  of neurons. These functions can change as the network's operating point does; for instance, as the leak reversal  $V_L$  increases, neurons will shift to a mean-driven, more oscillatory spiking regime and  $C^0(s)$  and  $K(t; A)$  will change. As the mean synaptic weight changes, the firing rates will change and this can also affect  $C^0(s)$  and  $K(t; A)$ . We have assumed weak synapses, so we will fix these at their value at  $p = p_0 W^{\max}/2$ . Without this approximation, the system is transcendental.

Each dynamical equation now contains four different sums of products of the weight and adjacency matrices. First examining  $dp/dt$ , we see that the first three sums correspond to defined motifs:  $1/N^2 \sum_{i,j} \mathbf{W}_{ij}^0 = p_0$ ,  $1/N^2 \sum_{i,j} \mathbf{W}_{ij}^0 \mathbf{W}_{ij} = p$  and  $1/N^2 \sum_{i,j} \mathbf{W}_{ij}^0 \mathbf{W}_{ji} = q_X^{\text{rec}} + pp_0$ . The last term in Eq. (32), however, corresponds to a third-order motif mixed between the weight and adjacency matrices. Similarly, third- and fourth-order mixed motifs appear in Eqs. 33 and 34. In order to calculate these, we extend a re-summing technique developed in [16]. We assume that there are no third- or higher-order correlations between elements of the weight and/or adjacency matrices, and approximate the frequency of each of these higher-order motifs by the number of ways it can be composed of one and two-synapse motifs. For a third order motif, this corresponds to adding up the likelihoods that all three synapses occur by chance and that each possible combination of one synapse and a two-synapse motif occur. In Eq. (32),

$$\sum_{i,j,k} \mathbf{W}_{ij}^0 \mathbf{W}_{ik} \mathbf{W}_{jk} \approx \epsilon^2 N^3 \left( p_0 (q^{\text{div}} + p^2) + p (q_X^{\text{con}} + q_X^{\text{ch,B}}) \right). \quad (39)$$

and for the four-synapse motif in Eq. (33),

$$\sum_{i,j,k,l} \mathbf{W}_{ik} \mathbf{W}_{jk}^0 \mathbf{W}_{jl} \mathbf{W}_{kl} \approx \epsilon^3 N^4 \left( p^3 p_0 + p^2 (q_X^{\text{div}} + q_X^{\text{con}} + q_X^{\text{ch,B}}) + pp_0 (q^{\text{div}} + q^{\text{ch}}) + q^{\text{div}} q_X^{\text{div}} + q^{\text{ch}} q_X^{\text{con}} \right) \quad (40)$$

This re-summing, along with the inclusion of the mixed motifs  $\{q_X\}$ , is what allows us to close the motif dynamics. Re-summing each third- and fourth-order motif in our system in terms of two-synapse motifs yields, after simplification, the final motif dynamics:

$$\frac{dp}{dt} = p_0 r^2 S + \epsilon \left[ p S_F + (q_X^{\text{rec}} + p_0 p) S_B + \frac{1}{p_0} \left( p_0 (q^{\text{div}} + p^2) + p (q_X^{\text{con}} + q_X^{\text{ch,B}}) \right) S_C \right] \quad (41)$$

$$\frac{dq^{\text{div}}}{dt} = 2r^2 S q_X^{\text{div}} + 2\epsilon \left[ q^{\text{div}} S_F + (p_0 q^{\text{ch}} + p q_X^{\text{div}}) S_B + \frac{1}{p_0} (q^{\text{ch}} (q_X^{\text{con}} + pp_0) + q_X^{\text{div}} (q^{\text{div}} + p^2)) S_C \right] \quad (42)$$

$$\frac{dq^{\text{con}}}{dt} = 2r^2 S q_X^{\text{con}} + 2\epsilon \left[ q^{\text{con}} S_F + (p_0 q^{\text{ch}} + p q_X^{\text{con}}) S_B + \frac{1}{p_0} \left( q^{\text{con}} (q_X^{\text{ch,B}} + p p_0) + q_X^{\text{con}} (q^{\text{div}} + p^2) \right) S_C \right] \quad (43)$$

$$\begin{aligned} \frac{dq^{\text{ch}}}{dt} = & r^2 S \left( q_X^{\text{ch,A}} + q_X^{\text{ch,B}} \right) + \epsilon \left[ 2q^{\text{ch}} S_F + \left( p_0 (q^{\text{con}} + q^{\text{div}}) + p (q_X^{\text{ch,A}} + q_X^{\text{ch,B}}) \right) S_B \right. \\ & \left. + \frac{1}{p_0} \left( q_X^{\text{ch,A}} (q^{\text{div}} + p^2) + q_X^{\text{ch,B}} (q^{\text{div}} + q^{\text{con}} + p^2) \right) S_C \right] \end{aligned} \quad (44)$$

$$\begin{aligned} \frac{dq_X^{\text{rec}}}{dt} = & r^2 S q_0^{\text{rec}} + \epsilon \left[ q_X^{\text{rec}} S_F + (1 - p_0) (q_X^{\text{rec}} + p p_0) S_B \right. \\ & \left. + \frac{1}{p_0} \left( q_0^{\text{rec}} (q^{\text{div}} + p^2) + q_X^{\text{ch,B}} (q_X^{\text{ch,B}} + p p_0) + q_X^{\text{con}} (q_X^{\text{con}} + p p_0) \right) S_C \right] \end{aligned} \quad (45)$$

$$\frac{dq_X^{\text{div}}}{dt} = r^2 S q_0^{\text{div}} + \epsilon \left[ q_X^{\text{div}} S_F + \left( p q_0^{\text{div}} + p_0 q_X^{\text{ch,B}} \right) S_B + \frac{1}{p_0} \left( q_0^{\text{div}} (q^{\text{div}} + p^2) + q_X^{\text{ch,B}} (q_X^{\text{con}} + p p_0) \right) S_C \right] \quad (46)$$

$$\frac{dq_X^{\text{con}}}{dt} = r^2 S q_0^{\text{con}} + \epsilon \left[ q_X^{\text{con}} S_F + \left( p q_0^{\text{con}} + p_0 q_X^{\text{ch,A}} \right) S_B + \frac{1}{p_0} \left( q_0^{\text{con}} (q^{\text{div}} + p^2) + q_X^{\text{con}} (q_X^{\text{ch,B}} + p p_0) \right) S_C \right] \quad (47)$$

$$\frac{dq_X^{\text{ch,A}}}{dt} = r^2 S q_0^{\text{ch}} + \epsilon \left[ q_X^{\text{ch,A}} S_F + (p q_0^{\text{ch}} + p_0 q_X^{\text{con}}) S_B + \frac{1}{p_0} \left( q_0^{\text{ch}} (q^{\text{div}} + p^2) + q_X^{\text{con}} (q_X^{\text{con}} + p p_0) \right) S_C \right] \quad (48)$$

$$\frac{dq_X^{\text{ch,B}}}{dt} = r^2 S q_0^{\text{ch}} + \epsilon \left[ q_X^{\text{ch,B}} S_F + (p q_0^{\text{ch}} + p_0 q_X^{\text{div}}) S_B + \frac{1}{p_0} \left( q_0^{\text{ch}} (q^{\text{div}} + p^2) + q_X^{\text{ch,B}} (q_X^{\text{ch,B}} + p p_0) \right) S_C \right] \quad (49)$$

Examination of these equations reveals how different types of joint spiking activity affect motif dynamics. Chance spiking coincidence (the  $r^2 S$  terms) couple each motif to the mixed version of itself, and each mixed motif to the baseline structure of the adjacency matrix. With Hebbian STDP and excitatory synapses,  $S_F > 0$  and  $S_B < 0$ . So, spiking covariance from forward connections provide positive feedback, reinforcing the current network structure. Spiking covariance from backward connections and common input couple divergent, convergent and chain motifs to each other.

The dynamics on the invariant set (Results: Balanced STDP of the mean synaptic weight, Fig. 6) were plotted in MATLAB. The vector fields of Figs. 8 and 9 were calculated in XPPAUT. For those figures, results from simulations of the full spiking network were plotted in MATLAB and then overlaid on the vector fields from XPPAUT.

### Plasticity of loops and feedforward chains

The chain variable  $q^{\text{ch}}$  includes both feedforward and recurrent loops. (Feedforward chains correspond to  $k \neq i$  in the definition of  $q^{\text{ch}}$ , Eq. (26), and recurrent loops to  $k = i$ .) As in the main text, we break  $q^{\text{ch}}$  into these two cases:  $q^{\text{ch}} = q^{\text{rec}} + q^{\text{ff}}$ , where

$$\begin{aligned} \epsilon^2 q^{\text{rec}} &= \frac{1}{N^3} \sum_{i,j,k} \delta_{ik} \mathbf{w}_{ij} \mathbf{w}_{jk} = \frac{1}{N^3} \sum_{i,j} \mathbf{w}_{ij} \mathbf{w}_{ji} \\ \epsilon^2 q^{\text{ff}} &= \frac{1}{N^3} \sum_{i,j,k} (1 - \delta_{ik}) \mathbf{w}_{ij} \mathbf{w}_{jk} - \epsilon^2 p^2 \end{aligned} \quad (50)$$

We also define an auxiliary variable which we will require in the dynamics of  $q^{\text{rec}}$ :

$$\epsilon^2 q_{X2}^{\text{rec}} = \frac{1}{N^3} \sum_{i,j} \mathbf{w}_{ij}^2 \mathbf{w}_{ji}^0 \quad (51)$$

which is proportional to the conditioned second moment of weights that are part of disynaptic loops. The dynamics of  $q^{\text{rec}}$  are calculated exactly as for the other motifs and are:

$$\frac{1}{2\epsilon} \frac{dq^{\text{rec}}}{dt} = r^2 S p_0 (q_X^{\text{rec}} + p p_0) + S_F q^{\text{rec}} + S_B q_{X2}^{\text{rec}} \quad (52)$$

where the new auxiliary variable obeys

$$\frac{1}{2\epsilon} \frac{dq_{X2}^{\text{rec}}}{dt} = r^2 S p_0 (q_X^{\text{rec}} + p p_0) + S_F q_{X2}^{\text{rec}} + S_B q^{\text{rec}} \quad (53)$$

We can then recover the dynamics of feedforward chains as:

$$\begin{aligned} \frac{dq^{\text{ff}}}{dt} &= \frac{dq^{\text{ch}}}{dt} - \frac{dq^{\text{rec}}}{dt} \\ &= r^2 S \left( q_X^{\text{ch,A}} + q_X^{\text{ch,B}} \right) + \epsilon \left[ -2r^2 S p_0 (q_X^{\text{rec}} + p p_0) + 2S_F (q^{\text{ch}} - q^{\text{rec}}) \right. \\ &\quad \left. + \left( p_0 (q^{\text{con}} + q^{\text{div}}) + p \left( q_X^{\text{ch,A}} + q_X^{\text{ch,B}} \right) - 2q_{X2}^{\text{rec}} \right) S_B \right. \\ &\quad \left. + \frac{1}{p_0} \left( q_X^{\text{ch,A}} (q^{\text{div}} + p^2) + q_X^{\text{ch,B}} (q^{\text{div}} + q^{\text{con}} + p^2) \right) S_C \right] \end{aligned} \quad (54)$$

### Unbalanced STDP

When there is an imbalance between the net amounts of potentiation and depression in the STDP rule, the motif dynamics are governed by simpler equations. If  $S \sim \mathcal{O}(1)$ , the  $\mathcal{O}(\epsilon)$  terms in Eqs. 41-49 are negligible. For each mixed motif,

$$q_X(t) = r^2 S q_0 t + q_X(0) \quad (55)$$

so that

$$p(t) = p_0 r^2 S t + p(0) \quad (56)$$

$$q^{\text{div}}(t) = q^{\text{div}}(0) + q_X^{\text{div}}(0) r^2 S t + \frac{1}{2} q_0^{\text{div}} (r^2 S)^2 t^2 \quad (57)$$

$$q^{\text{con}}(t) = q^{\text{con}}(0) + q_X^{\text{con}}(0) r^2 S t + \frac{1}{2} q_0^{\text{con}} (r^2 S)^2 t^2 \quad (58)$$

$$q^{\text{ch}}(t) = q^{\text{ch}}(0) + \left( q_X^{\text{ch,A}}(0) + q_X^{\text{ch,B}}(0) \right) r^2 S t + \frac{1}{2} q_0^{\text{ch}} (r^2 S)^2 t^2 \quad (59)$$

Writing  $q_X^{\text{ch}} = q_X^{\text{ch,A}} + q_X^{\text{ch,B}}$  puts the dynamics for all the motifs in the same form. The motifs expand from the initial conditions and baseline structure of the network. Note that since the quadratic term is proportional to  $S^2$ , even when STDP is depression-dominated the long-term dynamics are expansive rather than contractive.

## Acknowledgments

We thank Krešimir Josić and the members of the Doiron group for useful comments on the manuscript. Funding is provided by NSF-DMS1313225 (B.D.).

## References

1. Bullmore E, Sporns O (2009) Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience* 10: 186–198.
2. Markram H (1997) A network of tufted layer 5 pyramidal neurons. *Cerebral Cortex* 7: 523–533.
3. Perin R, Berger TK, Markram H (2011) A synaptic organizing principle for cortical neuronal groups. *Proceedings of the National Academy of Sciences* 108: 5419–5424.
4. Song S, Sjöström PJ, Reigl M, Nelson S, Chklovskii DB (2005) Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biol* 3: e68.
5. Yoshimura Y, Dantzker JLM, Callaway EM (2005) Excitatory cortical neurons form fine-scale functional networks. *Nature* 433: 868–873.
6. Ko H, Hofer SB, Pichler B, Buchanan KA, Sjöström PJ, et al. (2011) Functional specificity of local synaptic connections in neocortical networks. *Nature* 473: 87–91.
7. Yassin L, Benedetti BL, Jouhanneau JS, Wen JA, Poulet JFA, et al. (2010) An embedded subnetwork of highly active neurons in the neocortex. *Neuron* 68: 1043–1050.
8. Zhao L, Beverlin BI, Netoff T, Nykamp DQ (2011) Synchronization from second order network connectivity statistics. *Frontiers in Computational Neuroscience* 5: 28.
9. Roxin A (2011) The role of degree distribution in shaping the dynamics in networks of sparsely connected spiking neurons. *Frontiers in Computational Neuroscience* 5: 8.
10. Litwin-Kumar A, Doiron B (2012) Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nature Neuroscience* 15: 1498–1505.
11. Gaiteri C, Rubin JE (2011) The interaction of intrinsic dynamics and network topology in determining network burst synchrony. *Frontiers in Computational Neuroscience* 5: 10.
12. Kriener B, Helias M, Aertsen A, Rotter S (2009) Correlations in spiking neuronal networks with distance dependent connections. *Journal of Computational Neuroscience* 27: 177–200.
13. Pernice V, Staude B, Cardanobile S, Rotter S (2011) How structure determines correlations in neuronal networks. *PLoS Comput Biol* 7: e1002059.
14. Pernice V, Deger M, Cardanobile S, Rotter S (2013) The relevance of network micro-structure for neural dynamics. *Frontiers in Computational Neuroscience* 7.
15. Trousdale J, Hu Y, Shea-Brown E, Josić K (2012) Impact of network structure and cellular response on spike time correlations. *PLoS Computational Biology* 8: e1002408.
16. Hu Y, Trousdale J, Josić K, Shea-Brown E (2013) Motif statistics and spike correlations in neuronal networks. *Journal of Statistical Mechanics: Theory and Experiment* 2013: P03012.
17. Helias M, Tetzlaff T, Diesmann M (2014) The correlation structure of local neuronal networks intrinsically results from recurrent dynamics. *PLoS Comput Biol* 10: e1003428.
18. Hu Y, Trousdale J, Josić K, Shea-Brown E (2014) Local paths to global coherence: cutting networks down to size. *Physical Review E* : 032802.

19. Bi Gq, Poo Mm (1998) Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *The Journal of Neuroscience* 18: 10464–10472.
20. Markram H (1997) Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* 275: 213–215.
21. Abbott LF, Nelson SB (2000) Synaptic plasticity: taming the beast. *Nature neuroscience* 3: 1178–1183.
22. Caporale N, Dan Y (2008) Spike timing-dependent plasticity: A hebbian learning rule. *Annual Review of Neuroscience* 31: 25–46.
23. Markram H, Gerstner W, Sjöström PJ (2011) A history of spike-timing-dependent plasticity. *Frontiers in Synaptic Neuroscience* 3.
24. Song S, Miller KD, Abbott LF (2000) Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience* 3: 919–926.
25. Harris KD (2005) Neural signatures of cell assembly organization. *Nature Reviews Neuroscience* 6: 399–407.
26. Hebb DO (1949) *The organization of behavior: a neuropsychological theory*. Mahwah, N.J.: L. Erlbaum Associates.
27. Buzsáki G (2010) Neural syntax: cell assemblies, synapsembles, and readers. *Neuron* 68: 362–385.
28. Harris KD, Mrsic-Flogel TD (2013) Cortical connectivity and sensory coding. *Nature* 503: 51–58.
29. Kempter R, Gerstner W, Van Hemmen JL (2001) Intrinsic stabilization of output rates by spike-based hebbian learning. *Neural Computation* 13: 2709–2741.
30. Babadi B, Abbott LF (2010) Intrinsic stability of temporally shifted spike-timing dependent plasticity. *PLoS Comput Biol* 6: e1000961.
31. Gütig R, Aharonov R, Rotter S, Sompolinsky H (2003) Learning input correlations through nonlinear temporally asymmetric hebbian plasticity. *The Journal of neuroscience* 23: 3697–3714.
32. Rubin J, Lee D, Sompolinsky H (2001) Equilibrium properties of temporally asymmetric hebbian plasticity. *Physical Review Letters* 86: 364–367.
33. Kempter R, Gerstner W, Van Hemmen JL (1999) Hebbian learning and spiking neurons. *Physical Review E* 59: 4498.
34. Meffin H, Besson J, Burkitt AN, Grayden DB (2006) Learning the structure of correlated synaptic subgroups using stable and competitive spike-timing-dependent plasticity. *Physical Review E* 73: 041911.
35. Gilson M, Burkitt AN, Grayden DB, Thomas DA, Hemmen JL (2009) Emergence of network structure due to spike-timing-dependent plasticity in recurrent neuronal networks. i. input selectivity–strengthening correlated input pathways. *Biological Cybernetics* 101: 81–102.
36. Fiete IR, Senn W, Wang CZH, Hahnloser RHR (2010) Spike-time-dependent plasticity and heterosynaptic competition organize networks to produce long scale-free sequences of neural activity. *Neuron* 65: 563–576.



37. Kumar A, Rotter S, Aertsen A (2010) Spiking activity propagation in neuronal networks: reconciling different perspectives on neural coding. *Nat Rev Neurosci* 11: 615–627.
38. Izhikevich EM, Gally JA, Edelman GM (2004) Spike-timing dynamics of neuronal groups. *Cerebral Cortex* 14: 933–944.
39. Morrison A, Aertsen A, Diesmann M (2007) Spike-timing-dependent plasticity in balanced random networks. *Neural Computation* 19: 1437–1467.
40. Babadi B, Abbott LF (2013) Pairwise analysis can account for network structures arising from spike-timing dependent plasticity. *PLoS Computational Biology* 9: e1002906.
41. Litwin-Kumar A, Doiron B (2014) Formation and maintenance of neuronal assemblies through synaptic plasticity. *Nature Communications* 5.
42. Karbowski J, Ermentrout G (2002) Synchrony arising from a balanced synaptic plasticity in a network of heterogeneous neural oscillators. *Physical Review E* 65.
43. Burkitt AN, Gilson M, Hemmen JL (2007) Spike-timing-dependent plasticity for neurons with recurrent connections. *Biological Cybernetics* 96: 533–546.
44. Gilson M, Burkitt AN, Grayden DB, Thomas DA, Hemmen JL (2009) Emergence of network structure due to spike-timing-dependent plasticity in recurrent neuronal networks. II. input selectivity—symmetry breaking. *Biological Cybernetics* 101: 103–114.
45. Gilson M, Burkitt AN, Grayden DB, Thomas DA, Hemmen JL (2009) Emergence of network structure due to spike-timing-dependent plasticity in recurrent neuronal networks III: Partially connected neurons driven by spontaneous activity. *Biological Cybernetics* 101: 411–426.
46. Gilson M, Burkitt AN, Grayden DB, Thomas DA, Hemmen JL (2010) Emergence of network structure due to spike-timing-dependent plasticity in recurrent neuronal networks v: self-organization schemes and weight dependence. *Biological Cybernetics* 103: 365–386.
47. Doiron B, Lindner B, Longtin A, Maler L, Bastian J (2004) Oscillatory activity in electrosensory neurons increases with the spatial correlation of the stochastic input stimulus. *Phys Rev Lett* 93.
48. Lindner B, Doiron B, Longtin A (2005) Theory of oscillatory firing induced by spatially correlated noise and delayed inhibitory feedback. *Phys Rev E* 72.
49. Gerstner W, Kempter R, van Hemmen JL, Wagner H (1996) A neuronal learning rule for sub-millisecond temporal coding. *Nature* 383: 76–78.
50. Fourcaud-Trocme N, Hansel D, van Vreeswijk C, Brunel N (2003) How spike generation mechanisms determine the neuronal response to fluctuating inputs. *Journal of Neuroscience* 23: 11628–11640.
51. Jolivet R, Lewis TJ, Gerstner W (2004) Generalized integrate-and-fire models of neuronal activity approximate spike trains of a detailed model to a high degree of accuracy. *Journal of Neurophysiology* 92: 959–976.
52. Jolivet R, Schürmann F, Berger TK, Naud R, Gerstner W, et al. (2008) The quantitative single-neuron modeling competition. *Biological Cybernetics* 99: 417–426.
53. Gardiner C (2009) *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer Berlin Heidelberg.

54. Gilson M, Burkitt AN, Grayden DB, Thomas DA, Hemmen JL (2009) Emergence of network structure due to spike-timing-dependent plasticity in recurrent neuronal networks IV: Structuring synaptic pathways among recurrent connections. *Biological Cybernetics* 101: 427–444.
55. Feldman DE (2012) The spike-timing dependence of plasticity. *Neuron* 75: 556–571.
56. Froemke RC, Dan Y (2002) Spike-timing-dependent synaptic modification induced by natural spike trains. *Nature* 416: 433–438.
57. Kozloski J, Cecchi GA (2010) A theory of loop formation and elimination by spike timing-dependent plasticity. *Frontiers in Neural Circuits* 4.
58. Kunkel S, Diesmann M, Morrison A (2011) Limits to the development of feed-forward structures in large recurrent neuronal networks. *Frontiers in Computational Neuroscience* 4.
59. Galan R, Fourcaud-Trocme N, Ermentrout B, Urban NN (2006) Correlation-induced synchronization of oscillations in olfactory bulb neurons. *J Neurosci* 26: 3646–3655.
60. de la Rocha J, Doiron B, Shea-Brown E, Josić K, Reyes A (2007) Correlation between neural spike trains increases with firing rate. *Nature* 448: 802–6.
61. Shea-Brown E, Josić K, de la Rocha J, Doiron B (2008) Correlation and synchrony transfer in integrate-and-fire neurons: basic properties and consequences for coding. *Phys Rev Lett* 100.
62. Hong S, Ratte S, Prescott SA, De Schutter E (2012) Single neuron firing properties impact correlation-based population coding. *J Neurosci* 32: 1413–1428.
63. Ocker GK, Doiron B (2014) Kv7 channels regulate pairwise spiking covariability in health and disease. *Journal of Neurophysiology* 112: 340–352.
64. Deger M, Schwalger T, Naud R, Gerstner W (2013) Dynamics of interacting finite-sized networks of spiking neurons with adaptation. *arXiv:13114206 [q-bio]* .
65. Renart A, de la Rocha J, Bartho P, Hollender L, Parga N, et al. (2010) The asynchronous state in cortical circuits. *Science* 327: 587–590.
66. Brunel N (2000) Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *Journal of Computational Neuroscience* 8: 183–208.
67. Litwin-Kumar A, Chacron MJ, Doiron B (2012) The spatial structure of stimuli shapes the timescale of correlations in population spiking activity. *PLoS Comput Biol* .
68. Litwin-Kumar A, Oswald AMM, Urban NN, Doiron B (2011) Balanced synaptic input shapes the correlation between neural spike trains. *PLoS Comput Biol* 7: e1002305.
69. Sjöström PJ, Turrigiano GG, Nelson SB (2001) Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron* 32: 1149–1164.
70. Bi GQ, Wang HX (2002) Temporal asymmetry in spike timing-dependent synaptic plasticity. *Physiology & Behavior* 77: 551–555.
71. Wang HX, Gerkin RC, Nauen DW, Bi GQ (2005) Coactivation and timing-dependent integration of synaptic potentiation and depression. *Nature Neuroscience* 8: 187–193.
72. Wittenberg GM, Wang SSH (2006) Malleability of spike-timing-dependent plasticity at the CA3–CA1 synapse. *The Journal of Neuroscience* 26: 6610–6617.

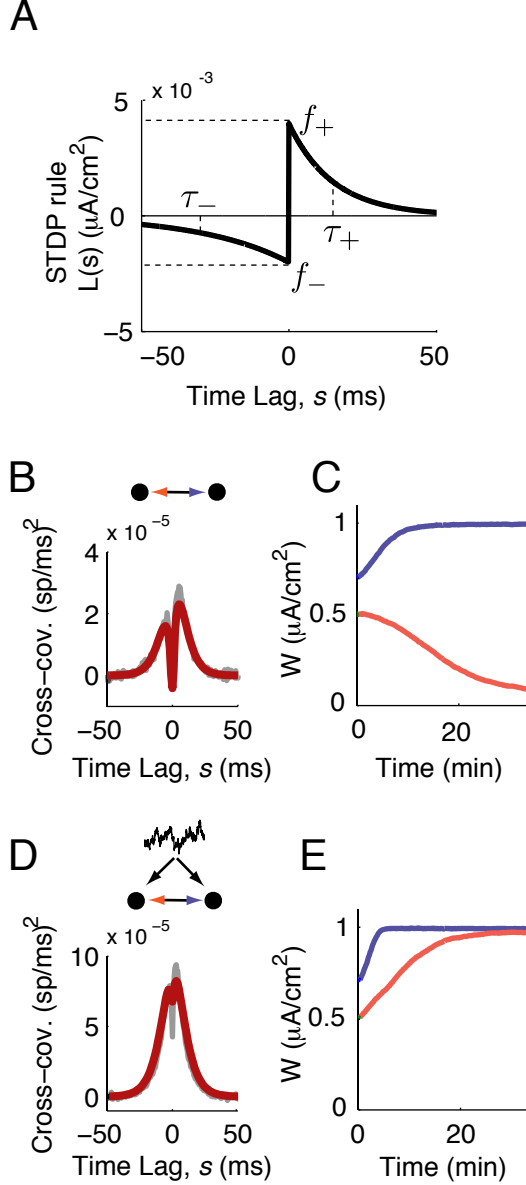
73. Pfister JP, Gerstner W (2006) Triplets of spikes in a model of spike timing-dependent plasticity. *The Journal of Neuroscience* 26: 9673–9682.
74. Clopath, Claudia, Büsing, Lars, Vasilaki, Eleni, Gerstner, Wulfram (2010) Connectivity reflects coding: a model of voltage-based STDP with homeostasis. *Nat Neurosci* 13: 344–352.
75. Shouval HZ, Bear MF, Cooper LN (2002) A unified model of NMDA receptor-dependent bidirectional synaptic plasticity. *Proceedings of the National Academy of Sciences of the United States of America* 99: 10831–10836.
76. Rubin JE, Gerkin, RC, Bi, G-Q, Chow, C (2005) Calcium time course as a signal for spike-timing-dependent plasticity. *Journal of Neurophysiology* 93: 2600–2613.
77. Graupner M, Brunel N (2012) Calcium-based plasticity model explains sensitivity of synaptic changes to spike pattern, rate, and dendritic location. *Proceedings of the National Academy of Sciences* 109: 3991–3996.
78. Churchland AK, Kiani R, Chaudhuri R, Wang XJ, Pouget A, et al. (2011) Variance as a signature of neural computations during decision making. *Neuron* 69: 818–31.
79. Kohn A, Smith MA (2005) Stimulus dependence of neuronal correlation in primary visual cortex of the macaque. *J Neurosci* 25: 3661–3673.
80. Churchland MM, Yu BM, Cunningham JP, Sugrue LP, Cohen MR, et al. (2010) Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature Neuroscience* 13: 369–378.
81. Arieli A, Sterkin A, Grinvald A, Aertsen A (1996) Dynamics of ongoing activity: Explanation of the large variability in evoked cortical responses. *Science* 273: 1868–1871.
82. Tsodyks M, Kenet T, Grinvald A, Arieli A (1999) Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science* 286: 1943–1946.
83. Lefort S, Tómm C, Sarria JCF, Petersen CCH (2009) The excitatory neuronal network of the c2 barrel column in mouse primary somatosensory cortex. *Neuron* 61: 301–316.
84. Ikegaya Y, Sasaki T, Ishikawa D, Honma N, Tao K, et al. (2013) Interpyramid spike transmission stabilizes the sparseness of recurrent network activity. *Cerebral Cortex* 23: 293–304.
85. Rubin JE (2001) Steady states in an iterative model for multiplicative spike-timing-dependent plasticity. *Network: Computation in Neural Systems* 12: 131–140.
86. Gilson M, Fukai T (2011) Stability versus neuronal specialization for STDP: Long-tail weight distributions solve the dilemma. *PLoS ONE* 6: e25339.
87. Royer S, Paré D (2003) Conservation of total synaptic weight through balanced synaptic depression and potentiation. *Nature* 422: 518–522.
88. Renart A, Song P, Wang XJ (2003) Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. *Neuron* 38: 473–485.
89. Lazar A, Pipa G, Triesch J (2009) SORN: a self-organizing recurrent neural network. *Frontiers in Computational Neuroscience* 3: 23.
90. Zheng P, Dimitrakakis C, Triesch J (2013) Network self-organization explains the statistics and dynamics of synaptic connection strengths in cortex. *PLoS Comput Biol* 9: e1002848.

91. Zenke F, Hennequin G, Gerstner W (2013) Synaptic plasticity in neural networks needs homeostasis with a fast rate detector. *PLoS Comput Biol* 9: e1003330.
92. Abraham WC (2008) Metaplasticity: tuning synapses and networks for plasticity. *Nature Reviews Neuroscience* 9: 387–387.
93. Levy N, Horn D, Meilijson I, Ruppin E (2001) Distributed synchrony in a cell assembly of spiking neurons. *Neural Networks: The Official Journal of the International Neural Network Society* 14: 815–824.
94. Mongillo G, Curti E, Romani S, Amit DJ (2005) Learning in realistic networks of spiking neurons and spike-driven plastic synapses. *The European Journal of Neuroscience* 21: 3143–3160.
95. Liu JK, Buonomano DV (2009) Embedding multiple trajectories in simulated recurrent neural networks in a self-organizing manner. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 29: 13172–13181.
96. Chen CC, Jasnow D (2010) Mean-field theory of a plastic network of integrate-and-fire neurons. *Physical Review E* 81: 011907.
97. Mayer J, Ngo HVV, Schuster HG (2012) Dynamical mean-field equations for a neural network with spike timing dependent plasticity. *Journal of Statistical Physics* 148: 677–686.
98. Ko H, Cossell L, Baragli C, Antolik J, Clopath C, et al. (2013) The emergence of functional microcircuits in visual cortex. *Nature* 496: 96–100.
99. Ko H, Mrsic-Flogel TD, Hofer SB (2014) Emergence of feature-specific connectivity of cortical microcircuits in the absence of visual experience. *J Neurosci* 34: 9812–9816.
100. Richardson M (2007) Firing-rate response of linear and nonlinear integrate-and-fire neurons to modulated current-based and conductance-based synaptic drive. *Phys Rev E* 76: 021919.
101. Richardson M (2008) Spike-train spectra and network response functions for non-linear integrate-and-fire neurons. *Biol Cybern* 99: 381–392.
102. Cox D, Isham V (1980) Point Processes. *Monographs on Statistics and Applied Probability*. CRC Press.

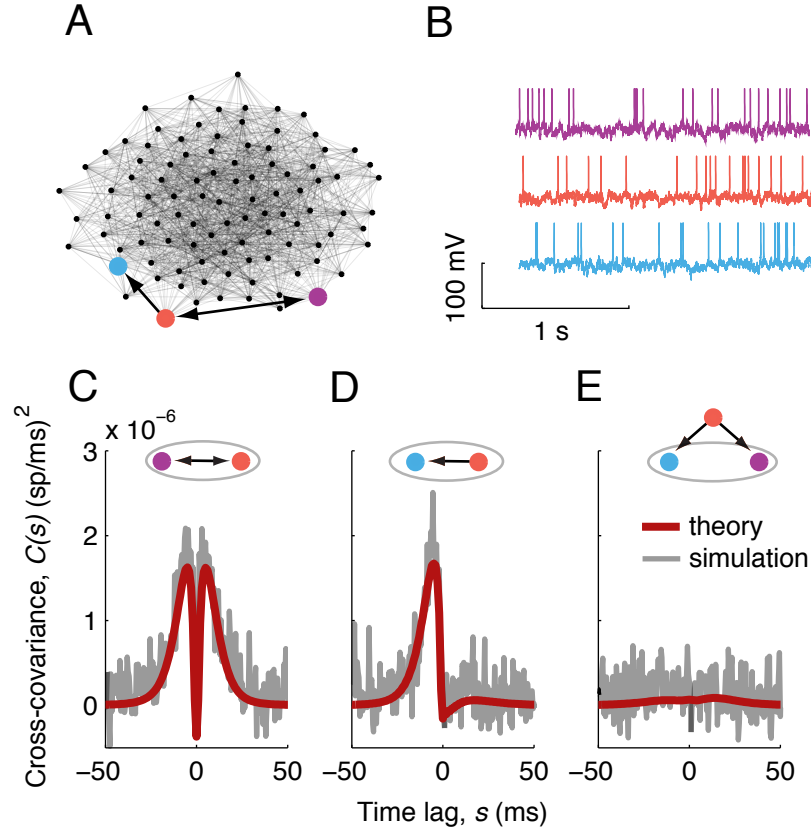
## Tables

**Table 1. Model parameters**

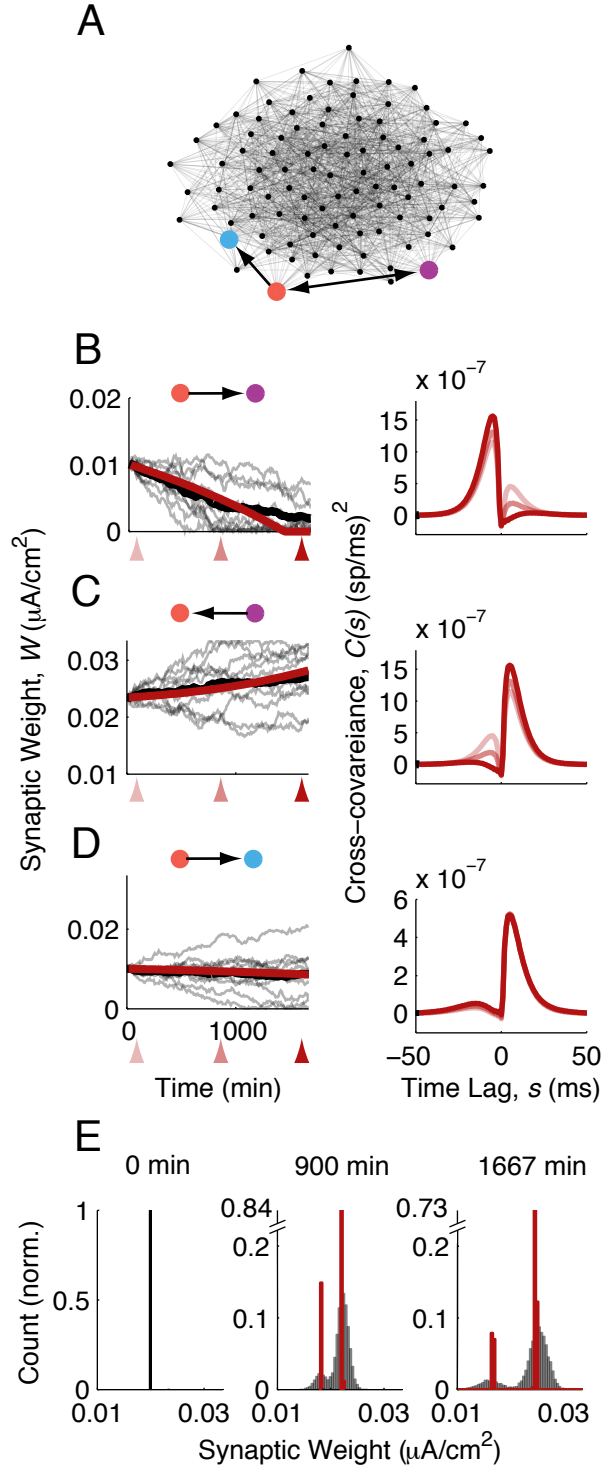
Parameter	Description	Value
$C$	Membrane capacitance	$1 \mu\text{F}/\text{cm}^2$
$g_L$	Leak conductance	$0.1\text{mS}/\text{cm}^2$
$V_L$	Leak reversal potential	-72 mV
$\Delta$	Action potential steepness	1.4 mV
$V_T$	Action potential initiation threshold	-48 mV
$V_{th}$	Action potential threshold	30 mV
$V_{re}$	Action potential reset	-72 mV
$\tau_{ref}$	Action potential width	2 ms
$\mu$	External input mean	$1 \mu\text{A}/\text{cm}^2$
$\sigma$	External input standard deviation	9 mV
$N$	Number of neurons	1000
$p_0$	Connection density	.15
$W^{\max}$	Maximum synaptic weight	$5 \mu\text{A}/\text{cm}^2$
$\tau_S$	Synaptic time constant	5 ms



**Figure 1. Network structure shapes synaptic plasticity.** (A) The STDP rule,  $L(s)$ , is composed of exponential windows for depression (-) and potentiation (+). Each is defined by its amplitude  $f_{\pm}$  and timescale  $\tau_{\pm}$ . (B) Spike train cross-covariance function for a pair of neurons with no common input, so that synapses between the two neurons are the only source of spiking covariance. Shaded lines: simulation, solid lines: theory (Eq. (4)). (C,E) Synaptic weight (peak EPSC amplitude) as a function of time in the absence (C) and presence (E) of common input. (D) Spike train cross-covariance function for a pair of neurons with common input,  $c = 0.05$ . Common input was modeled as an Ornstein-Uhlenbeck process with a 5 ms timescale.

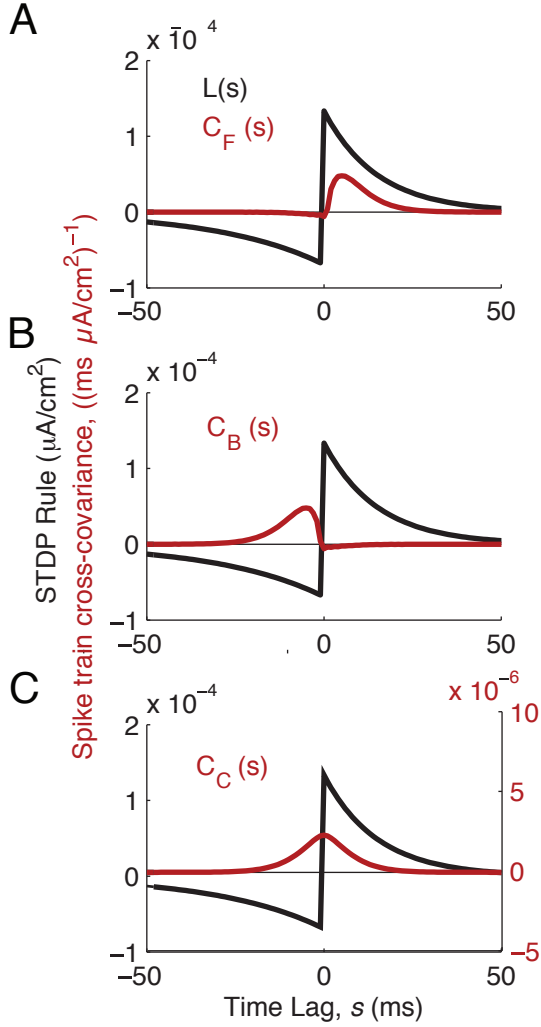


**Figure 2. Linear response theory for spiking covariances.** (A) Illustration of the network connectivity for a subset of 100 neurons. Three neurons, and the connections between them, are highlighted. Nodes are positioned by the Fruchterman-Reingold force algorithm. (B) Example voltage traces for the three highlighted neurons. (C-E) Spike train cross-covariance functions for the three combinations of labeled neurons. Top: A shaded ellipse contains the pair of neurons whose cross-covariance is shown. Shaded lines: simulations, red lines: linear response theory.

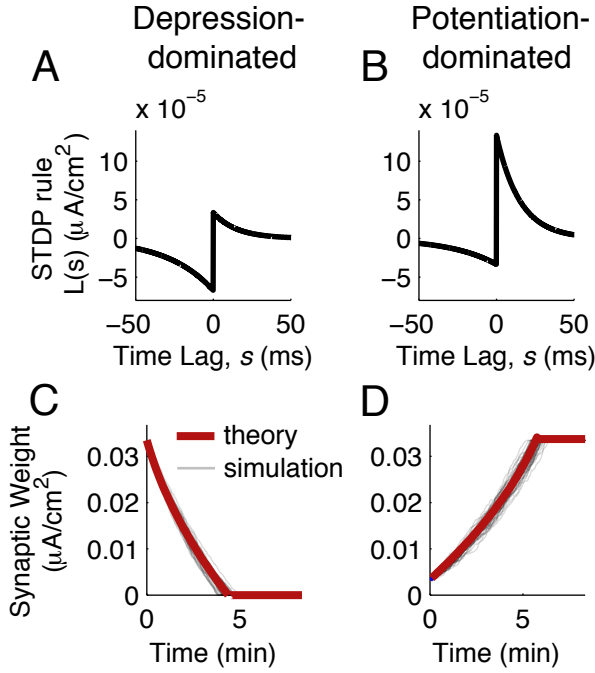


**Figure 3. STDP in recurrent networks with internally generated spiking covariance.** (A) As in Fig. 2A. (B-D) Left, Synaptic weight versus time for each of the three synapses in the highlighted network. Shaded lines: simulation, individual trials. Solid black lines: simulation, trial-average. Solid red lines: theory. Right, spike train cross-covariances at the three time points marked on the left (linear response theory). (E) Histogram of synaptic weights at three time points. Red, theory. Shaded: simulation.

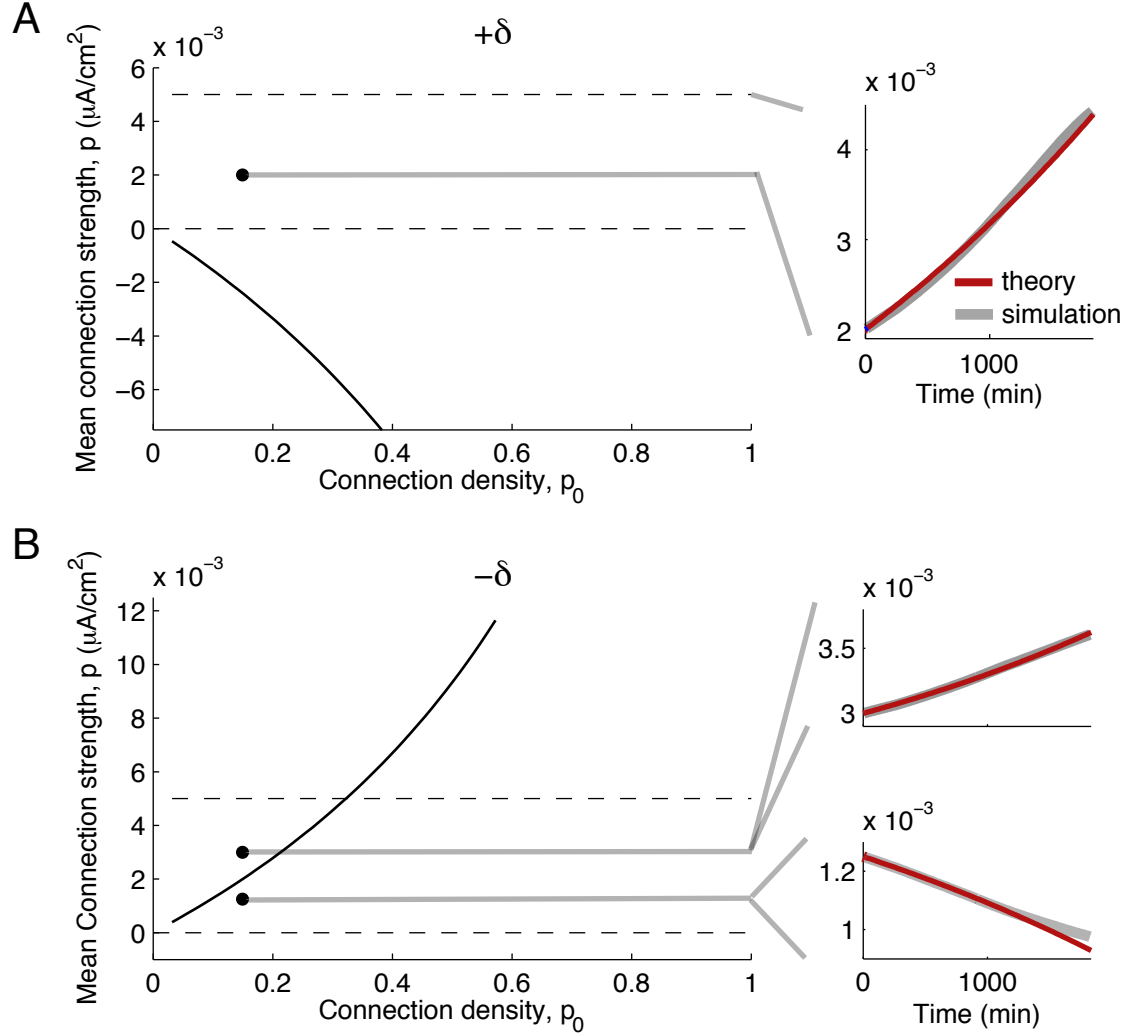




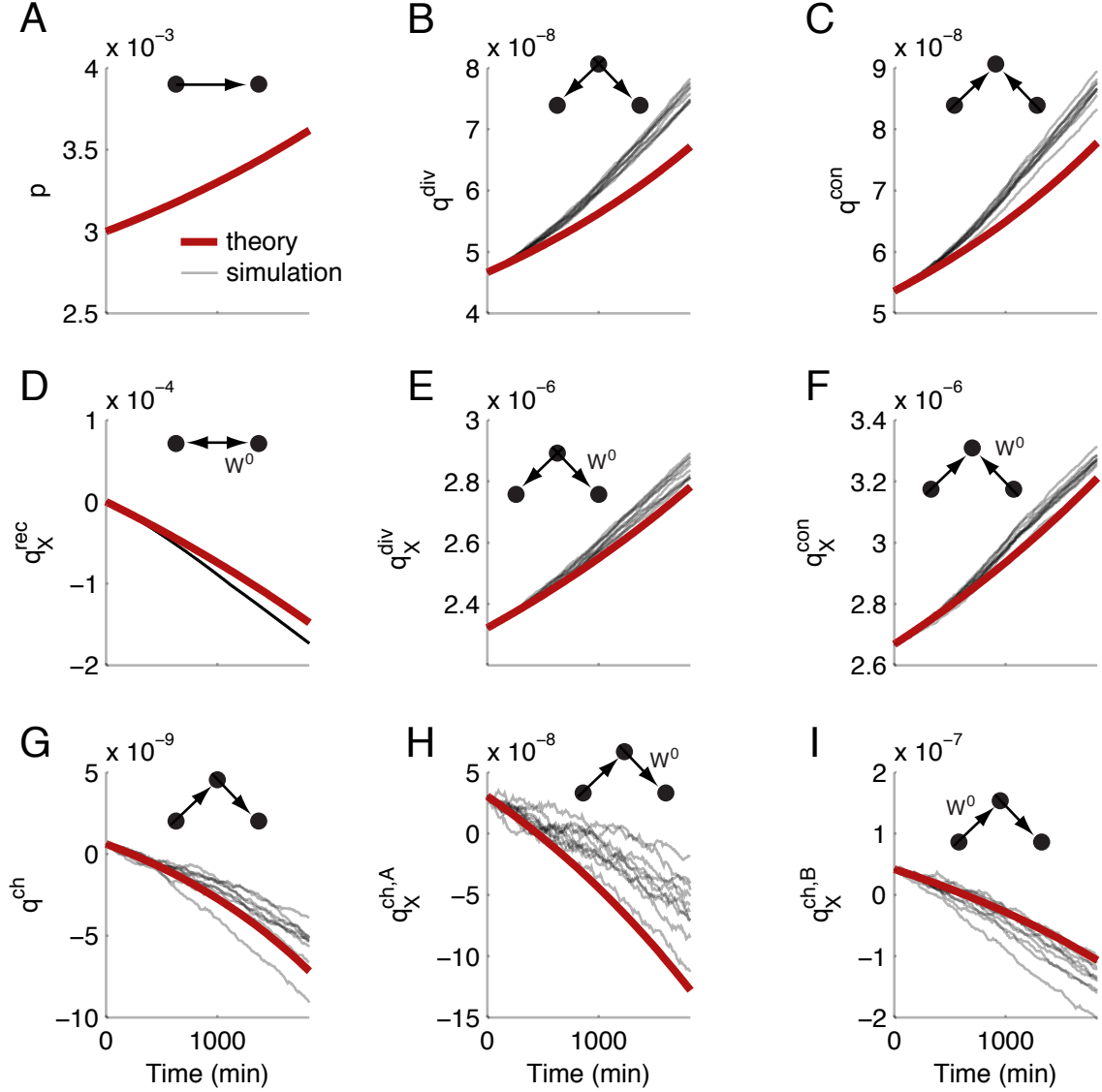
**Figure 4. Different sources of spiking covariance interact with different parts of the STDP rule.** Black: STDP rule. Red: spike train cross-covariances, from Eq. (7). (A) Covariance from forward connections interacts with the potentiation side of the STDP rule. (B) Covariance from backward connections interacts with the depression side of the STDP rule. (C) Covariance from common input is temporally symmetric and interacts with both the potentiation and depression sides of the STDP rule.



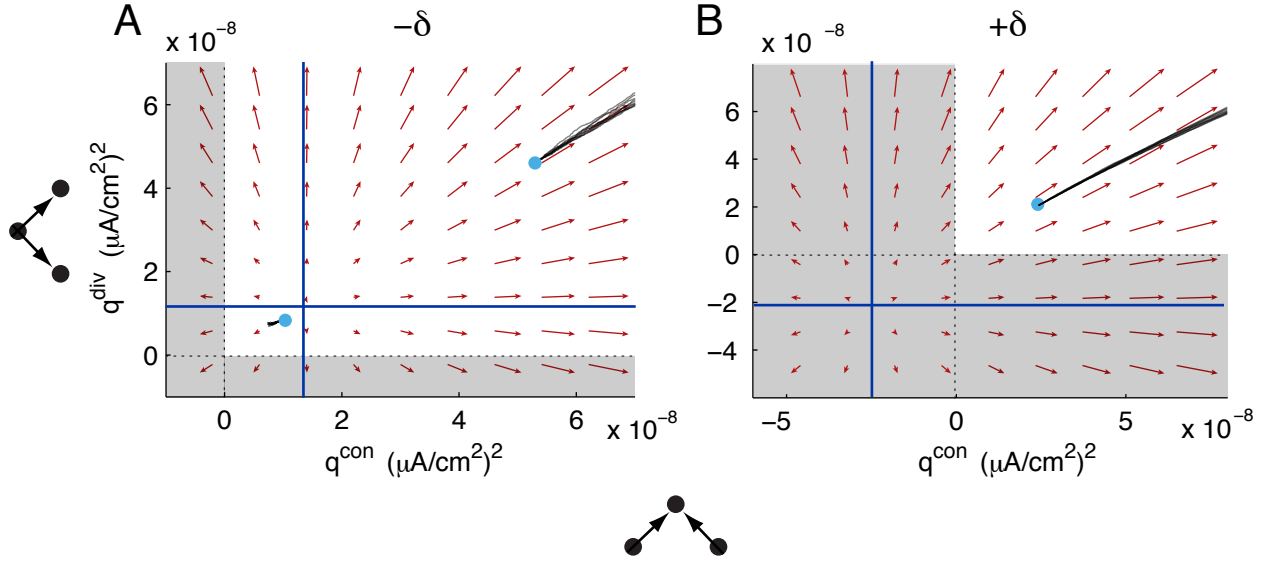
**Figure 5. Unbalanced plasticity gives rise to simple weight dynamics.** (A) A depression-dominated STDP rule: the amount of depression (integral of the depression side of the curve) is twice the amount of potentiation. (B) A potentiation-dominated STDP rule: the amount of potentiation is twice the amount of depression. (C) Evolution of synaptic weights with depression-dominated STDP: all weights depress. (D) Evolution of synaptic weights with potentiation-dominated STDP: all weights potentiate. Red lines: theory for mean synaptic weight. Shaded lines: simulation of individual synaptic weights.



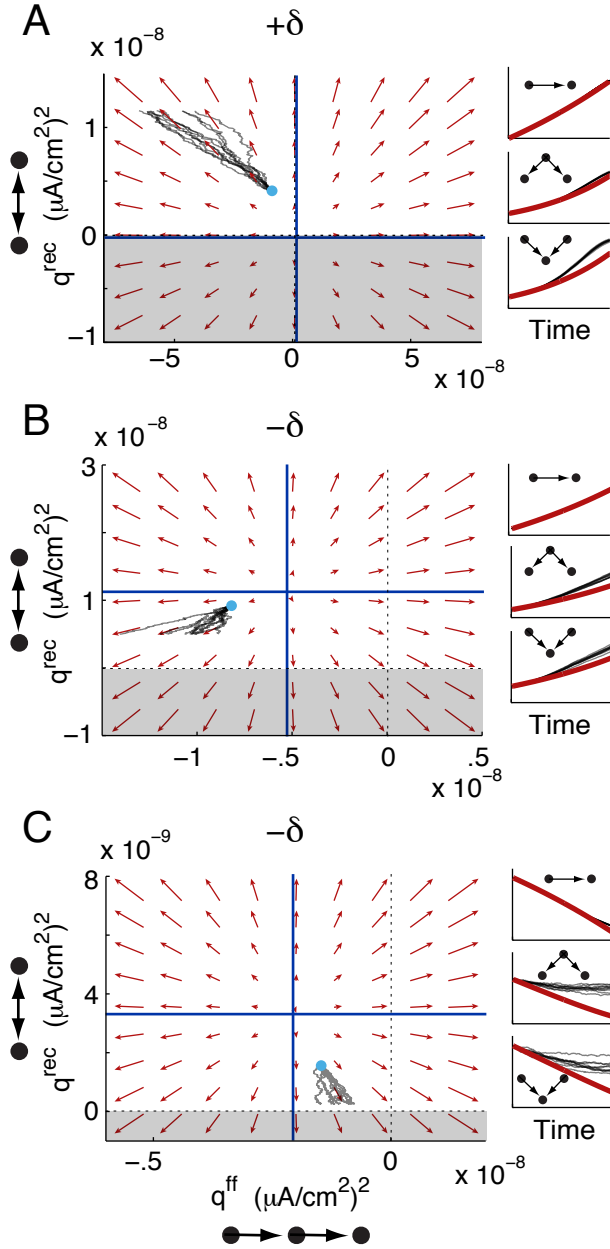
**Figure 6. Balanced plasticity of the mean synaptic weight.** (A) When the STDP rule is balanced and potentiation-dominated, the unstable fixed point for  $p$  is negative and decreases with the connection probability. So, the mean synaptic weight always increases. (B) When the STDP rule is balanced and depression-dominated, the unstable fixed point is positive and increases with the connection probability. (A,B) Left: Dashed lines mark bounds for the mean synaptic weight, at 0 and  $p_0 W^{\max}$ . Black curves track the location of the unstable fixed point of  $p$  as the connection probability,  $p_0$ , varies. Black dots mark initial conditions for the right panels. (A,B) Right: Dynamics of the mean synaptic weight in each of the regimes of the left plots. Red lines mark the reduced theory's prediction (Eq. (9)) and shaded lines the result of simulating the full spiking network.



**Figure 7. Reduced theory for the plasticity of two-synapse motifs.** In each panel, the strength of a different motif or mixed motif is plotted as it evolves. (A) Mean synaptic weight. (B) Divergent motifs. (C) Convergent motifs. (D) Mixed recurrent motifs (strength of connections conditioned on their being part of a two-synapse loop). (E) Mixed divergent motifs (strength of individual synapses conditioned on their being part of a divergent motif). (F) Mixed convergent motifs. (G) Chain motifs. (H) Mixed chains type A (strength of individual synapses conditioned on their being the first in a chain). (I) Mixed chains type B (strength of individual synapses conditioned on their being the second in a chain). The STDP rule was in the depression-dominated balanced regime, as in Fig. 7B.



**Figure 8. Plasticity of convergent and divergent motifs with balanced STDP.** (A) Joint dynamics of convergent and divergent motifs when STDP is balanced and depression-dominated. Initial conditions as in Fig. 7A. (B) Joint dynamics of convergent and divergent motifs when STDP is balanced and potentiation-dominated. Initial conditions as in Fig. 7B. Red: the flow in the  $q^{\text{con}}, q^{\text{div}}$  slice of the motif phase space. Black: plasticity of the motifs in simulations of the full spiking network. Cyan dots mark initial conditions for the simulations and each line is a realization. The vector fields are calculated with all other variables fixed at these initial conditions. For (A), the vector fields are similar for both sets of initial conditions. In both panels, blue lines mark projections of each variable's nullcline into the plane and regions of unattainable negative motif strengths are shaded.



**Figure 9. Plasticity of recurrent loops and feedforward chains with balanced STDP.** (A-C) The dynamics of loops and feedforward chains, with all other variables fixed at their initial conditions. In all cases, the projections of the  $q^{\text{ff}}$  and  $q^{\text{rec}}$  nullclines into this plane provide thresholds for the potentiation or depression of each motif. The shape of the STDP rule and the initial values of the other motif variables determine the locations of these nullclines. Color conventions are as in Fig. 8. In each panel, right insets show the time series of  $p$  (top),  $q^{\text{div}}$  (middle) and  $q^{\text{con}}$  (bottom), with spiking simulations in black and motif theory in red. A) The potentiation-dominated balanced STDP rule. B) The depression-dominated balanced STDP rule, in the region where  $p$ ,  $q^{\text{div}}$  and  $q^{\text{con}}$  potentiate. C) The depression-dominated balanced STDP rule, in the region where  $p$ ,  $q^{\text{div}}$  and  $q^{\text{con}}$  depress.